FACULTEIT ECONOMIE EN
BEDRIJFSWETENSCHAPPEN

KU LEUVEN

Sparse estimation of high-dimensional time series models

Proefschrift Voorgedragen tot
het Behalen van de Graad van
Doctor in de Toegepaste
Economische Wetenschappen

door

Ines WILMS

NUMMER 522 2016

Committee

Advisor:	
Prof. Dr. Christophe Croux	*KU Leuven*
Chair:	
Prof. Dr. Robert Boute	*KU Leuven*
Members:	
Prof. Dr. Geert Dhaene	*KU Leuven*
Prof. Dr. Sarah Gelper	*Technische Universiteit Eindhoven*
Prof. Dr. Jeroen Rombouts	*ESSEC Business School*
Prof. Dr. Martina Vandebroek	*KU Leuven*

Acknowledgments

This thesis concludes almost four years of intensive research. Several people have contributed to this work by giving professional and personal support.

First and foremost, I would like to express profound gratitude to my supervisor Prof. Christophe Croux. Thank you for sharing your expertise, for always believing in me, for providing a stimulating and motivating research environment, and for passing on your passion for conducting research. Conducting research and communicating it in an understandable way is not an easy task. Also here, thank you for being an excellent 'teacher'. All chapters of this thesis have greatly improved by your suggestions and valuable feedback. It has been a privilege and pleasure to work with you and I hope we will continue our collaborations.

I would also like to thank all members of the doctoral committee: Prof. Geert Dhaene, Prof. Sarah Gelper, Prof. Jeroen Rombouts, and Prof. Martina Vandebroek. Thank you for the thorough reading of my thesis, and for all constructive comments during the doctoral seminars and preliminary defense. Your comments have led to a considerable improved quality of this text. Sarah, our collaborations on Chapter 1 and 3 have been very instructive, motivating and inspiring to me.

Furthermore, I am very grateful to the FWO (Fonds Wetenschappelijk Onderzoek Vlaanderen) for the financial support and to KU Leuven for providing me with the facilities to carry out this doctoral research. The statistical seminars organized by the LSTAT group have definitely broadened my knowledge.

I would also like to thank all my colleagues from ORSTAT, Heverlee and the 'fourth floor'. Thank you for the pleasant working atmosphere. I really enjoyed our numerous joint activities. Special thanks to Viktoria for providing support from the beginning of my PhD until the end.

Tot slot wil ik nog graag mijn familie bedanken. Bedankt om er altijd voor mij te zijn, om me in alles aan te moedigen en te steunen.

Ines Wilms Leuven, April 2016.

Introduction

Nowadays, a large amount of data is available in nearly every area of science and business. Information is typically collected in data sets where the different variables are contained in the columns of the data set and the measurements on each variable are contained in the rows. Our interest mainly lies in settings where these measurements are collected over time. Such data sets are said to contain *time series* in their columns. A time series should be treated differently from a regular variable to account for the time-dependency of its measurements. As an example, data sets in marketing containing weekly sales, price and promotional information on product categories (e.g. soft drinks) are typically available.

Moreover, given today's data abundance, our interest lies in *high-dimensional* data sets, as opposed to *low-dimensional* data sets. High-dimensional time series data sets contain many short time series: a large number of time series (columns) is available relative to the number of time points (rows), hence, these data sets are 'fat'. Low-dimensional time series data sets, in contrast, contain few long time series: a large number of time points (rows) is available relative to the number of time series (columns), hence, these data sets are 'thin'. High-dimensional time series data sets are commonplace in today's business practice since many firms collect information on a large number of variables, but discard data that are older than a few years. For instance, consider the goal in marketing to predict future sales volumes for a large number of product categories based on their past sales, price and promotional information. To obtain such predictions, an estimation method is needed.

The problem, however, is that traditional estimators are well suited for low-dimensional data sets, but not for high-dimensional data sets. On the one hand, these estimators suffer from very low estimation precision if the number of measurements (rows) is close to the number of variables (columns) in the data set. This leads towards inaccurate predictions. On the other hand, traditional estima-

tors are not even computable if the number of measurements (rows) in the data set is larger than the number of variables (columns). Then, no predictions can be made. Hence, there is a need for new estimation methods especially designed for these high-dimensional data sets.

In this thesis, we develop *sparse* estimation methods for high-dimensional data. Despite the data abundance, we do not expect each variable of these data sets to be equally informative. Sparse estimation methods rely on a simplicity assumption: we assume that only a relative small number of variables in our data set plays an important role. As such, sparse estimators retain the informative variables and remove the non-informative ones. In our marketing example, we do not expect that each variable will influence each category's sales volume. Instead, we expect some variables to be unimportant in predicting these sales volumes, but we do not know which. Sparse estimators detect for each category the variables that are important for predicting its future sales volume, and it estimates the effect these variables have on its future sales volume. For the variables that are detected to be unimportant, their estimated effect is put to zero. This highly facilitates interpretation.

We develop sparse estimators for high-dimensional time series models in Chapters 1 to 4, and for Canonical Correlation Analysis (CCA) in Chapters 5 and 6. CCA is a multivariate statistical method that describes the associations between two data sets. Our interest lies in settings where both data sets are high-dimensional. Throughout the thesis, the usefulness and relevance of the sparse estimators are discussed for a wide variety of application domains, ranging from marketing (Chapter 1), and economics (Chapter 3, 4), to biometrics (Chapter 2, 5, 6).

Chapter 1 is dedicated to the development of a sparse estimator for the high-dimensional Vector AutoRegressive (VAR) model. The VAR model is the preferred model in econometrics to analyze the relationship between several time series. We focus on high-dimensional VAR models, models containing a large number of short time series. We show that more accurate estimation and prediction results are obtained with our sparse estimator compared to traditional estimators. We apply the sparse estimator of the VAR to a high-dimensional marketing data set where we predict sales volumes for a large number of product categories. The sparse estimator yields insightful results regarding which categories are more influential (meaning that they are important drivers of other category's sales), and which categories are more responsive (meaning that they react to changes in other categories).

In Chapter 2, we develop a computational algorithm for sparse estimation of the general class of Multivariate Regression Models, of which the VAR model is a special case. In high-dimensional Multivariate Regression Models, a large number of response variables (not necessarily time series) is predicted using a large number of predictor variables. We illustrate the algorithm on a biometric data set. Biometrics is another field of science where high-dimensional data sets are commonplace. We consider a genomic data set that contains information on a large number of genes, but only a small number of measurements is available (often because of cost constraints).

Chapter 3 considers the popular time series concept of 'Granger Causality' in high-dimensions. A (set of) time series is said to Granger Cause another time series if the former has incremental, or additional, predictive power for the latter. We develop a new test procedure, called the 'Granger Lasso test', to test for Granger Causality in high-dimensional settings. We show that this test is more powerful than traditional tests in such settings. We use the proposed test to study the predictive power of economic sentiment indicators for future macro-economic developments. We find that forecast accuracy is improved by using only the most predictive sentiment indicators, obtained with the Granger Lasso test, rather than all indicators.

Chapter 4 considers another popular time series concept in high-dimensions, namely 'Cointegration'. Cointegration analysis is used to estimate long-run equilibrium relations between several time series. The coefficients of these relations are called the 'cointegrating vectors'. We provide a sparse estimator of the cointegrating vectors. The proposed estimator is used for interest rate growth forecasting and consumption growth forecasting. We show that it leads to important gains in forecast accuracy compared to traditional estimators.

In Chapter 5, we develop a sparse CCA method. The typical research field of interest is biometrics where one wants to study associations between one data set containing gene expression data and another containing comparative genomic hybridization data. Identifying associations between both is extremely important to increase our understanding of the development of diseases such as cancer. A sparse approach is often wanted to identify the most important variables for the association study. We illustrate the good performance of our proposed sparse CCA method compared to other sparse CCA alternatives by means of a simulation study and a biometric data example.

Finally, in Chapter 6, we address the frequent occurrence of outliers in high-dimensional data sets used for CCA. Outliers are observations with an atypical

behavior, making it unlikely that they are generated by the model. In genomics, some patients can react very differently to treatments because of their individual-specific genetic structure. The possible presence of outlying observations should be taken into account and estimates should remain reliable, or 'robust', in their presence. In Chapter 6, we therefore robustify the sparse CCA method from Chapter 5. An additional advantage of the proposed robust sparse CCA method is that outliers can be identified. Knowledge of such atypical patients is extremely useful for geneticists.

The various chapters in this thesis can be found in

(i) S. Gelper, I. Wilms and C. Croux. Identifying demand effects in a large network of product categories. *Journal of Retailing*, 92(1), 25-39, 2016.

(ii) I. Wilms and C. Croux. An algorithm for the multivariate group lasso with covariance estimation. *FEB Research Report* KBI_1528, 2015.

(iii) I. Wilms, S. Gelper and C. Croux. The predictive power of the business and bank sentiment of firms: A high-dimensional Granger Causality approach. *European Journal of Operational Research,* Accepted, 2016.

(iv) I. Wilms and C. Croux. Forecasting using sparse cointegration. *International Journal of Forecasting*, Accepted, 2016.

(v) I. Wilms and C. Croux. Sparse canonical correlation analysis from a predictive point of view. *Biometrical Journal*, 57(5), 834-851, 2015.

(vi) I. Wilms and C. Croux. Robust sparse canonical correlation analysis. *FEB Research Report* KBI_1428, 2014.

Table of contents

Chapter 1

Identifying demand effects in a large network of product categories

Abstract

Planning marketing mix strategies requires retailers to understand within- as well as cross-category demand effects. Most retailers carry products in a large variety of categories, leading to a high number of such demand effects to be estimated. At the same time, we do not expect cross-category effects between all categories. This paper outlines a methodology to estimate a parsimonious product category network without prior constraints on its structure. To do so, sparse estimation of the Vector AutoRegressive Market Response Model is presented. We find that cross-category effects go beyond substitutes and complements, and that categories have asymmetric roles in the product category network. Destination categories are most influential for other product categories, while convenience and occasional categories are most responsive. Routine categories are moderately influential and moderately responsive.

1.1 Introduction

While within-category demand effects of the marketing mix have been studied extensively, cross-category effects are less well understood [Leeflang and Selva,

2012]. Nevertheless, cross-category effects might be substantial. Some categories are complements, e.g. bacon and eggs studied by Niraj et al. [2008] or cake mix and cake frosting studied by Manchanda et al. [1999], while others are substitutes, e.g. frozen, refrigerated and shelf-stable juices [Wedel and Zhang, 2004]. But cross-effects also exist among categories that are not complements or substitutes for several reasons. First, as a result of brand extensions, brands are no longer limited to one category [Erdem, 1998, Kamkura and Kang, 2007, Ma et al., 2012]. So advertising and promotion of a brand within one category might spill over to own brand sales in other categories. Second, advertising and promotions generate more store traffic and therefore more sales in other categories [Bell et al., 1998]. And third, lower expenditures in one category alleviate the budget constraint such that consumers are able to spend more on other, seemingly unrelated, categories [Song and Chintagunta, 2007, Lee et al., 2013].

While cross-category effects might be substantial for these reasons, we do not expect that each category's marketing mix variables influence each and every other category. Instead, we expect some cross-category effects to be zero – or very close to zero – but we can not a priori exclude them. Therefore, we use an exploratory modeling approach for parsimonious estimation of a product category network. The network allows us to easily identify categories that are influential for or responsive to changes in other categories. Building on a widely used category typology of destination, routine, occasional and convenience categories [Blattberg et al., 1995, Briesch et al., 2013], we find that destination categories are most influential, convenience and occasional categories most responsive, and routine categories moderately influential and moderately responsive.

In order to estimate the cross-category network, this paper presents sparse estimation of the Vector AutoRegressive (VAR) model. The estimation is *sparse* in the sense that some of the within-and cross-category effects in the model can be estimated as exactly zero. Initiated by the work of Baghestani [1991] and Dekimpe and Hanssens [1995], the VAR Market Response Model has become a standard, flexible tool to measure own- and cross-effects of marketing actions in a competitive environment. The main drawback of the VAR model is the risk of overparametrization because the number of parameters increases quadratically with the number of included categories. Earlier studies using the VAR model, like e.g. Nijs et al. (2001; 2007); Pauwels et al. [2002]; Srinivasan et al. (2000; 2004); Steenkamp et al. [2005], were often limited by this overparametrization problem. To overcome this problem, previous research on cross-category effects has limited its attention to a small number of categories by studying substitutes or comple-

ments [Kamkura and Kang, 2007, Song and Chintagunta, 2007, Leeflang et al., 2008, Bandyopadhyay, 2009, Ma et al., 2012]. We present an estimation technique for cross-category effects in much larger product category networks. The technique allows many parameters to be estimated even with short observation periods. Short observation periods are commonplace in marketing practice since many firms discard data that are older than one year [Lodish and Mela, 2007].

This paper contributes to the extant retail literature in a number of important ways. (1) Previous cross-category literature largely limits attention to categories that are directly related through substitution, complementarity or brand extensions. We provide evidence that cross-category effects go beyond such directly related categories. (2) We introduce the concepts of influence and responsiveness of a product category and position different category types (destination, routine, occasional and convenience) according to these dimensions. (3) To identify the cross-category effects, we estimate a large VAR model using an extension of the lasso approach of Tibshirani [1996].

The remainder of this article is organized as follows. Section 1.2 positions this paper in the cross-category management literature and describes the conceptual framework that positions category types according to their influence and responsiveness. Section 1.3 discusses the methodology. We describe the sparse estimator of the VAR model, discuss how to construct impulse response functions and compare the sparse estimation technique with two Bayesian estimators. In Section 1.4, a simulation study shows the excellent performance of the proposed methodology in terms of estimation reliability and prediction accuracy. Section 1.5 presents our data and model, Section 1.6 our findings on cross-category demand effects. We first identify which categories are most influential and which are most responsive to changes in other categories. Then, we identify the main cross-category effects based on estimated cross-price, promotion and sales elasticities.

1.2 Cross-Category Management

The importance of category management for retailers is widely acknowledged, both as a marketing tool for category performance [Fader and Lodish, 1990, Basuroy et al., 2001, Dhar et al., 2001] and as an operational tool for planning and logistics [Rajagopalan and Xia, 2012]. Successful category management requires retailers to understand cross-category effects of prices, promotions and sales. Among these, the cross-category effects of prices on sales – which define substitutes and complements – are the most extensively studied [Song and Chin-

tagunta, 2006, Bandyopadhyay, 2009, Leeflang and Selva, 2012, Sinistyn, 2012]. Cross-category effects of promotions, e.g. feature and display promotions, on sales result from many brands being active in multiple categories [Erdem and Sun, 2002]. Brand associations carry over to products of the same brand in other categories, e.g. through umbrella branding [Erdem, 1998] or horizontal product line extensions [Aaker and Keller, 1990]. Less well understood than the effects of prices and promotions, are the effects of sales in one category on sales in other categories. Such effects might exist because categories are related based on affinity in consumption [Shankar and Kannan, 2014], because products from various categories are placed close to each other in the shelves [Bezawada et al., 2009, Shankar and Kannan, 2014], or because of the budget constraint [Du and Kamakura, 2008]. If consumers spend more in a certain category they might, all else equal, spend less in other categories simply because they hit their budget constraint. As a result, cross-category effects might exist between seemingly unrelated categories.

When studying these cross-category effects of price, promotion and sales on sales, several asymmetries might arise. A first asymmetry concerns within- versus cross-category effects. We expect within-category effects to be more prevalent and larger in size than cross-category effects (e.g. Song and Chintagunta, 2006; Bezawada et al., 2009). A second asymmetry concerns category influence versus category responsiveness. Influential categories are important drivers of other category's sales, while sales of responsive categories react to changes in other categories. To identify which categories are more influential or more responsive, we build on a widely used typology of categories described in Blattberg et al. [1995].

Blattberg et al. [1995] define 4 category types from the consumer perspective: destination, routine, occasional and convenience. Destination categories contain goods that consumers plan to buy before they go on a shopping trip, such as soft drinks. Briesch et al. [2013] show that destination categories are generally categories in which consumers spend a lot of their budget. Retailers typically use a price aggressive promotion strategy and high promotion intensity for these destination categories with the goal of increasing store traffic. Because consumers shop to buy products in the destination categories, destination categories are likely to influence sales in other categories. However, since consumers already plan to buy in the destination categories before entering the store, destination category sales will not be highly responsive [Shankar and Kannan, 2014].

About 55% to 60% of categories are routine categories [Pradhan, 2009]. Routine categories are regularly and routinely purchased, such as juices and biscuits. Retailers typically use a consistent pricing strategy and average level of promo-

tion intensity. Because purchases in routine categories can more easily be delayed than purchases in destination categories, we expect routine categories to be more responsive. But, since purchases in routine categories altogether still account for a large portion of the budget, they are also likely to influence sales in other categories.

Occasional categories follow a seasonal pattern or are purchased infrequently. These categories comprise a small proportion of retail expenditures while they contain typically more expensive items, like oatmeal. We therefore expect occasional categories to be less influential and more responsive than destination or routine categories.

Finally, convenience categories are categories that consumers find convenient to pick up during their one-stop shopping trip, like ready-to-eat-meals. These purchase decisions are typically made in the store. Since convenience categories are geared towards consumer convenience and filling impulse needs, we expect them to be highly responsive.

1.3 Sparse Vector Auto-Regressive Modeling

1.3.1 Motivation

The aim of this paper is to identify cross-category demand effects in a large product category network. To this end, we use the Vector AutoRegressive (VAR) model. The VAR is ideal for measuring within- and cross-category effects of marketing actions since it accounts for both inertia in marketing spending and performance feedback effects by treating marketing variables as endogenous [Dekimpe and Hanssens, 1995]. Other studies on cross-category effects, like e.g. Wedel and Zhang [2004] use a demand model with exogenous prices, or a simultaneous equations model without lagged effects like Shankar and Kannan [2014]. However, managers may set marketing instruments strategically in response to market performance and market response expectations. Not accounting for time inertia or feedback effects limits our understanding of how the market functions and misleads managerial insights and prediction.

Identifying cross-category demand effects using VAR analysis remains challenging because the sheer number of such effects makes them hard to estimate. The number of parameters to be estimated in the VAR rapidly explodes, making standard estimation inaccurate. This undermines the ability to identify important relationships in the data. To overcome an explosion of the number of parameters

in the VAR, marketing researchers have used pre-estimation dimension reduction techniques, i.e. they first impose restrictions on the model and then estimate the reduced model. Four such common techniques are (i) treating marketing variables as exogenous (e.g. Nijs et al., 2001; Pauwels et al., 2002 and Nijs et al., 2007), (ii) estimating submodels rather than a full model (e.g. Srinivasan et al., 2000; Srinivasan et al., 2004), (iii) aggregating or pooling over, for instance, stores or competitors (e.g. Horvath et al., 2005; Slotegraaf and Pauwels, 2008), and (iv) applying Least Squares to a restricted model (e.g. Dekimpe and Hanssens, 1995, Dekimpe et al., 1999; Nijs et al., 2007). Most researchers applying pre-estimation dimension reduction techniques recognize that they do so because of the practical limitations of standard estimation techniques rather than for theoretical reasons (e.g. Srinivasan et al., 2004 and Bandyopadhyay, 2009).

To address the overparametrization of the VAR, we use sparse estimation. Sparsity means that some of the within- and cross-category effects in the VAR are estimated as exactly zero. As argued in the previous section, from a substantive perspective, we cannot exclude cross-category effects before estimation because cross-category effects might occur between seemingly unrelated categories. From a methodological perspective, sparse estimation is a powerful solution to handle the overparametrization of the VAR. In our cross-category model, we endogenously model sales, promotion and prices of 17 product categories. Hence, already in a VAR model with one lag, as much as $(3 \times 17) \times (3 \times 17) = 2601$ within- and cross-category effects need to be estimated. Since the sparse estimation procedure puts some of these effects to zero, a more parsimonious model is obtained. Results are easier to interpret and, therefore, the sparse estimation procedure provides actionable insights to managers.

1.3.2 Extending the Lasso to the VAR model

In situations where the number of parameters to estimate is large relative to the sample size, the Lasso proposed by Tibshirani [1996] provides a solution within the multiple regression model. The Lasso minimizes the least squares criterion penalized for the sum of the absolute values of the regression parameters. This penalization forces some of the estimated regression coefficients to be exactly zero, which results in selection of the pertinent variables in the model. The Lasso method is well established [Bühlmann and van de Geer, 2011, Chatterjee and Lahiri, 2011] and shows good performance in various applied fields [Wu et al., 2009, Fan et al., 2011].

The Lasso technique can not be directly applied to the VAR model because the VAR model differs from a multiple regression model in two important aspects. First, a VAR model contains several equations, corresponding to a multivariate regression model. Correlations between the error terms of the different equations need to be taken into account. Second, a VAR model is dynamic, containing lagged versions of the same time series as right-hand side variables of the regression equation. Both aspects of VAR models make it necessary to extend the lasso to the VAR context, what the sparse estimator in this paper does.

It builds further on a sparse estimator of the multivariate regression model [Rothman et al., 2010], and the groupwise lasso for categorical variables [Yuan and Lin, 2006, Meier et al., 2008]. The estimator is consistent for the unknown model parameters, see Meier et al. [2008] and Friedman et al. [2008].

1.3.3 Model Specification

Sales, price and promotion are measured for several categories over a certain time period. We collect all these time series in a multivariate time series $\mathbf{y}_t$ with q components. In our cross-category demand effects study, $\mathbf{y}_t$ contains sales, price and promotion for 17 product categories, hence $q = 3 \times 17 = 51$. The VAR Market Response Model is given by

$$\mathbf{y}_t = \mathbf{B}_1 \mathbf{y}_{t-1} + \mathbf{B}_2 \mathbf{y}_{t-2} + \ldots + \mathbf{B}_p \mathbf{y}_{t-p} + \mathbf{e}_t \,, \tag{1.1}$$

where p is the lag length. The autoregressive parameters $\mathbf{B}_1$ to $\mathbf{B}_p$ are $(q \times q)$ matrices, which capture both within- and cross-category effects. The elements of these matrices measure the effect of sales, price and promotion in one category on the sales, price and promotion in other categories (including its own). The error term $\mathbf{e}_t$ is assumed to follow a $N_q(\mathbf{0}, \mathbf{\Sigma})$ distribution. We assume, without loss of generality, that all time series are mean centered such that no intercept is included.

If the number of components q in the multivariate time series is large, the number of unknown elements in the sequence of matrices $\mathbf{B}_1, \ldots, \mathbf{B}_p$ explodes to pq^2, and accurate estimation by standard methods is no longer possible. Sparse estimation, with many elements of the matrices $\mathbf{B}_1, \ldots, \mathbf{B}_p$ estimated as zero, brings an outcome: it will not only provide estimates with smaller mean squared error, but also substantially improve model interpretability. The method we propose does not require the researcher to prespecify which entries in the $\mathbf{B}_j$ matrices are zero and which are not. Instead, the estimation and variable selection are si-

multaneously performed. This is particularly of interest in situations where there is no a priori information on which time series is driving which.

The instantaneous correlations in model (1.1) are captured in the error covariance matrix $\mathbf{\Sigma}$. If the dimension q is large relative to the number of observations, estimation of $\mathbf{\Sigma}$ becomes problematic. The estimated covariance matrix risks getting singular, i.e. its inverse does not exist. Hence, we also induce sparsity in the estimation of the inverse error covariance matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$. The elements of $\mathbf{\Omega}$ have a natural interpretation as partial correlations between the error components of the q equations in model (1.1). If the ij-th element of the inverse covariance matrix is zero this means that, conditional on the other error terms, there is no correlation between the error terms of equations i and j.

1.3.4 Penalized Likelihood Estimation

This section defines the sparse estimation procedure for the VAR model. The Sparse VAR estimator is defined by minimizing a measure of goodness-of-fit to the data combined with a *penalty* for the magnitude of the model parameters. It is convenient to first recast model (1.1) in stacked form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}\,, \tag{1.2}$$

where $\mathbf{y}$ is a vector of length nq containing the stacked values of the time series. If the multivariate time series has length T, then $n = T - p$ is the number of time points for which all current and lagged observations are available. The vector $\boldsymbol{\beta}$ contains the stacked vectorized matrices $\mathbf{B}_1, \ldots, \mathbf{B}_p$, and $\mathbf{e}$ the vector of stacked error terms. The matrix $\mathbf{X} = \mathbf{I}_q \otimes \mathbf{X}_0$, with $\mathbf{X}_0 = (\underline{\mathbf{Y}}_1, \ldots, \underline{\mathbf{Y}}_p)$, is of dimension $(nq \times pq^2)$. Here $\underline{\mathbf{Y}}_j$ is an $(n \times q)$ matrix, containing the values of the q series at lag j in its columns, for $1 \leq j \leq p$, with p the maximum lag. The symbol $\otimes$ stands for the Kronecker product.

The sparse estimator of the autoregressive parameters $\boldsymbol{\beta}$ and the inverse covariance matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ are obtained by minimizing the negative log likelihood with a groupwise penalization on the $\boldsymbol{\beta}$ and a penalization on the off-diagonal elements of $\mathbf{\Omega}$:

$$(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{\Omega}}) = \underset{(\boldsymbol{\beta}, \mathbf{\Omega})}{\operatorname{argmin}} \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \widetilde{\mathbf{\Omega}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \log|\mathbf{\Omega}| + \lambda_1 \sum_{g=1}^{G} ||\boldsymbol{\beta}_g|| + \lambda_2 \sum_{k \neq k'} |\Omega_{kk'}|\,, \tag{1.3}$$

where $||\mathbf{u}|| = (\sum_{i=1}^{n} u_i^2)^{1/2}$ is the Euclidean norm and $\widetilde{\mathbf{\Omega}} = \mathbf{\Omega} \otimes \mathbf{I}_n$. By simultaneously estimating $\boldsymbol{\beta}$ and $\mathbf{\Omega}$, we take the correlation structure between the error

terms into account. The vector $\boldsymbol{\beta}_g$ in (1.3) is a subvector of $\boldsymbol{\beta}$, containing the regression coefficients for the lagged values of the same time series in one of the q equations in model (1.1). The coefficients of the lagged values of the same time series form a group. The total number of groups is $G = q^2$ because there are q groups within each of the q equations. The penalty on the regression coefficients enforces that either *all* elements of the group $\widehat{\boldsymbol{\beta}}_g$ are zero or *none*. As a result, we take the dynamic nature of the VAR model into account since the estimated $\mathbf{B}_j$ matrices, for $j = 1, \ldots, p$, have their zero elements in exactly the same cells. The penalization on the off-diagonal elements of $\boldsymbol{\Omega}$ induces sparsity in the estimate $\widehat{\boldsymbol{\Omega}}$. Finally, the scalars λ_1 and λ_2 control the degree of sparsity of the regression estimator and the inverse covariance matrix estimator, respectively. The larger these values, the more sparsity is imposed. Details on the algorithm to perform penalized likelihood estimation and the selection of the sparsity parameters λ_1 and λ_2 can be found in Appendix 1.8.

Our approach is similar to Hsu et al. [2008] who use the Lasso within a VAR context. However, they do not account for the group-structure in the VAR model, nor do they impose sparsity on the error covariance matrix. Davis et al. [2015] propose another sparse estimation procedure for the VAR. They infer the sparsity structure of the autoregressive parameters from an estimate of the partial spectral coherence using a two-step procedure. Since variable selection is performed prior to model estimation, the resulting estimator suffers from pre-testing bias. Moreover, the number of parameters might still approach the sample size, leading to unstable estimation or even making estimation infeasible if the number of parameters still exceeds the sample size. Sparse estimation in economics is a growing field, see Fan et al. [2011] and references therein for an overview.

1.3.5 Alternative: Bayesian Estimators

An alternative to the sparse estimation technique is to impose prior information in a Bayesian setting. Bayesian regularization techniques have been proposed for the VAR model in Litterman [1980] and are used in various applied fields such as macroeconomics [Gefang, 2014, Banbura et al., 2010], finance [Carriero et al., 2012] and marketing [Lenk and Orme, 2009, Horvath and Fok, 2013, Bandyopadhyay, 2009]. They are also applicable to a situation like ours where there are many parameters to be estimated with a limited observation period, and are thus a good benchmark. However, these methods are not sparse, they do not perform variable selection simultaneously with model estimation. The following two paragraphs

elaborate on two Bayesian estimators which serve as non-sparse alternatives.

Minnesota Prior. The original Minnesota prior only specifies a prior distribution for the regression parameters of the VAR model. The error covariance matrix $\boldsymbol{\Sigma}$ is assumed to be diagonal, and estimated by $\hat{\Sigma}_{ii} = \hat{\sigma}_i^2$ with $\hat{\sigma}_i^2$ the standard OLS estimate of the error variance in an AR(p) model for the i^{th} time series [Koop and Korobilis, 2009]. The prior distribution of the regression parameters is taken to be multivariate normal:

$$\boldsymbol{\beta} \sim N(\underline{\boldsymbol{\beta}}_M, \underline{\mathbf{V}}_M). \tag{1.4}$$

For the prior mean, the common choice is $\underline{\boldsymbol{\beta}}_M = \mathbf{0}_{Kq}$ for stationary series. The prior covariance matrix $\underline{\mathbf{V}}_M$ is diagonal. The posterior distribution is again multivariate normal. Full technical details can be found in Koop and Korobilis [2009].

The main advantage of the Minnesota prior is its ease of implementation, since posterior inference only involves the multivariate normal distribution. However, imposing the Minnesota prior only ensures that the parameter estimates are *shrunken* towards zero, while the Sparse VAR ensures that some parameters will be estimated as *exactly* zero.

Normal Inverted Wishart Prior. The Minnesota prior takes the error covariance matrix $\boldsymbol{\Sigma}$ as fixed and diagonal and, hence, not as an unknown parameter. To overcome this problem, Banbura et al. [2010] impose an inverse Wishart prior on the $\boldsymbol{\Sigma}$ matrix. More precisely,

$$\boldsymbol{\beta} \mid \boldsymbol{\Sigma} \sim N(\underline{\boldsymbol{\beta}}_{NIW}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Omega}_0) \quad \text{and} \quad \boldsymbol{\Sigma} \sim iW(\mathbf{S}_0, \nu_0), \tag{1.5}$$

where $\underline{\boldsymbol{\beta}}_{NIW}, \boldsymbol{\Omega}_0, \mathbf{S}_0$ and ν_0 are hyperparameters. Under this normal inverted Wishart prior (labeled in the remainder of this paper as 'NIW'), the posterior for $\boldsymbol{\beta}$, conditional on $\boldsymbol{\Sigma}$ is normal, and the posterior for $\boldsymbol{\Sigma}$ is again inverted Wishart. Full technical details can be found in Banbura et al. [2010].

1.3.6 Impulse Response Functions

Impulse response functions (IRFs) are extensively used to assess the dynamic effect of external shocks to the system such as changes in the marketing mix. An IRF pictures how a change to a certain variable at moment t impacts the value of any other time series at time $t + k$, accounting for interrelations with all other variables. The magnitude of the effect is plotted as a function of k. An extensive discussion on the interpretation of the IRF in marketing modeling can

be found in Dekimpe and Hanssens [1995]. We use IRFs to gain insight in the dynamics of within and cross-category sales, promotion and price effects on each of the 17 product category sales. The IRFs are easily computed as a function of the Sparse VAR estimator (see Hamilton, 1991). Since we want to account for correlated error terms, we use generalized IRFs [Pesaran and Shin, 1998, Dekimpe and Hanssens, 1999].

To obtain confidence bounds for the generalized IRFs estimated by Sparse VAR, we use a residual parametric bootstrap procedure [Chatterjee and Lahiri, 2011]. We generate $N_b = 1000$ time series of length T from the VAR model (1.2). The invertible estimate of $\mathbf{\Sigma}$ delivered by the Sparse VAR estimation procedure is needed to draw random numbers for the $N_q(\mathbf{0}, \mathbf{\Sigma})$ error distribution. For each of these N_b multiple time series, the estimates of the regression parameters are computed. We compute the covariance matrix of the N_b bootstrap replicates. For each of the N_b generated series impulse response functions are computed; the 90% confidence bounds are then obtained by taking the 5% and 95% percentiles.

1.4 Estimation and prediction performance

We conduct a simulation study to compare the proposed Sparse VAR with Bayesian methods using the Minnesota and NIW prior. As benchmarks, we include the classical Least Squares (LS) estimator and two restricted versions of LS which are often used in practice. In the 1-step Restricted LS [Dekimpe and Hanssens, 1995, Dekimpe et al., 1999], we estimate the model with classical LS, delete all variables with $|t\text{-statistic}| \leq 1$, and re-estimate the model with the remaining variables. We also consider an iterative Restricted LS method described in Lütkepohl and Kratzig [2004] where we fit the full model using LS and sequentially eliminate the variables leading to the largest reduction of BIC until no further improvement is possible, of which a close variant was used by Nijs et al. [2007].

We simulate from a VAR model with $q = 30$ dimensions and $p = 2$ lags. Each time series has an own auto-regressive structure and we include system dynamics among the different series. The first series leads series 2 to 15, while the 16^{th} series leads time series 17 to 30. Specifically, the data generating processes are given by

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{B}_1 & 0 \\ 0 & \mathbf{B}_1 \end{bmatrix} \mathbf{y}_{t-1} + \begin{bmatrix} \mathbf{B}_2 & 0 \\ 0 & \mathbf{B}_2 \end{bmatrix} \mathbf{y}_{t-2} + \mathbf{e}_t \,,$$

with

$$\mathbf{B}_1 = \begin{bmatrix} \mathbf{0.4}_{1\times 1} & \mathbf{0}_{1\times 14} \\ \mathbf{0.4}_{14\times 1} & 0.4\cdot\mathbf{I}_{14} \end{bmatrix} \text{ and } \quad \mathbf{B}_2 = \begin{bmatrix} \mathbf{0.2}_{1\times 1} & \mathbf{0}_{1\times 14} \\ \mathbf{0.2}_{14\times 1} & 0.2\cdot\mathbf{I}_{14} \end{bmatrix}.$$

In total, there are $pq^2 = 1800$ regression parameters to be estimated with 116 true parameter values different from zero. The 30-dimensional error term $\mathbf{e}_t$ is drawn from a multivariate normal with mean zero and covariance matrix $\mathbf{\Sigma} = 0.1\mathbf{I}_{30}$. We generate $N_s = 1000$ multivariate time series of length 80 according to the above simulation scheme.

1.4.1 Performance measures

We evaluate the different estimators in terms of (i) estimation accuracy, (ii) sparsity recognition performance, and (iii) forecast performance.

To evaluate estimation accuracy, we compute the mean absolute estimation error (MAEE), averaged over the simulation runs and over the 1800 parameters

$$\text{MAEE} = \frac{1}{N_s}\frac{1}{pq^2}\sum_{s=1}^{N_s}\sum_{j=1}^{p}\sum_{k,l=1}^{q}|\hat{b}_{klj}^s - b_{klj}|,$$

where $\hat{b}_{klj}^s$ is the estimate of b_{klj}, the kl^{th} element of the matrix $\mathbf{B}_j$ corresponding to lag j, for the s^{th} simulation run.

Concerning sparsity recognition, we compute the true positive rate and true negative rate

$$\text{TPR} = \frac{1}{N_s}\sum_{s=1}^{N_s}\frac{\#\{(k,l,j) : \hat{b}_{klj}^s \neq 0 \ and \ b_{klj} \neq 0\}}{\#\{(k,l,j) : \ b_{klj} \neq 0\}}$$

$$\text{TNR} = \frac{1}{N_s}\sum_{s=1}^{N_s}\frac{\#\{(k,l,j) : \hat{b}_{klj}^s = 0 \ and \ b_{klj} = 0\}}{\#\{(k,l,j) : \ b_{klj} = 0\}}.$$

The true positive rate (TPR) gives an indication on the number of true relevant regression parameters detected by the estimation procedure. The true negative rate (TNR) measures the hit rate of detecting a true zero regression parameter. Both should be as large as possible.

Finally, we conduct an out-of-sample rolling window forecasting exercise. Using the same simulation design as before, we generate multivariate time series of length $T = 90$, and use a rolling window of length $S = 80$. For all estimation methods, 1-step-ahead forecasts are computed for $t = S, \ldots, T-1$. Next, we

Table 1.1: *Mean Absolute Estimation Error (MAEE), True Positive Rate (TPR), True Negative Rate (TNR) and Mean Absolute Forecast Error (MAFE), averaged over 1000 simulation runs, are reported for every method.*

Method	MAEE	TPR	TNR	MAFE
Sparse VAR	0.025	0.991	0.661	0.429
LS	0.189	1	0	1.069
Restricted LS: 1-step	0.159	0.638	0.390	0.800
Restricted LS: Iterative	0.114	0.643	0.444	0.779
Bayesian: Minnesota	0.027	1	0	0.427
Bayesian: NIW	0.061	1	0	0.628

compute the mean absolute forecast error (MAFE), averaged over all time series and across time

$$\text{MAFE} = \frac{1}{T-S}\frac{1}{q}\sum_{t=S}^{T-1}\sum_{i=1}^{q} \mid \hat{y}_{t+1}^{(i)} - y_{t+1}^{(i)} \mid,$$

where $y_{t+1}^{(i)}$ is the value of the i^{th} time series at time $t+1$.

1.4.2 Results

Table 1.1 presents the performance measures of the Sparse VAR, the Bayesian and benchmark methods. The Sparse VAR estimator performs best in terms of estimation accuracy. It attains the lowest value of the MAEE (0.025). A paired t-test confirms that the Sparse VAR significantly outperforms the other methods (all p-values < 0.01).

Sparsity recognition performance is evaluated using the true positive rate and the true negative rate, reported in Table 1.1. For the LS and Bayesian estimators, all parameters are estimated as non-zero, resulting in a perfect true positive rate and zero true negative rate. Among the variable selection methods, the Sparse VAR performs best. Sparse VAR achieves a value of the true positive rate of 0.99; 0.66 for the true negative rate.

Finally, we evaluate the forecast performance of the different estimators by the Mean Absolute Forecast Error in Table 1.1. The Sparse VAR and the Bayesian estimator with Minnesota prior achieve the best forecast performance. A Diebold-Mariano test [Diebold and Mariano, 1995] confirms that these two methods per-

form significantly better than the others (p-values < 0.01). There is no significant difference in forecast performance between Sparse VAR and the Bayesian estimator with Minnesota prior.

1.4.3 Robustness checks

Alternative penalty function. We investigate the robustness of Sparse VAR to the choice of the penalty function. We replace the grouplasso penalty on the regression coefficients with the elastic net penalty [Zou and Hastie, 2005]. Elastic net is a regularized regression method that linearly combines the L_1 and L_2 penalties of respectively lasso and ridge regression. Like the grouplasso, elastic net produces a sparse estimate of the regression coefficients. All other steps of the methodology remain unchanged. We find that the grouplasso penalty performs slightly better than the elastic net penalty in terms of estimation accuracy, sparsity recognition and prediction performance.

Sensitivity to the order of the VAR. We estimate the model with Sparse VAR for different values of p and evaluate the performance. As expected, Sparse VAR attains the best estimation accuracy for the true value $p = 2$. The results are, however, very robust to the choice of the order of the VAR. Selecting p too low is slightly worse than selecting p too high.

Sensitivity to the sparsity parameters. The sparsity parameters are selected according to the BIC and this selection is an integral part of the estimation procedure. The results are not sensitive to the value of λ_2, which controls the sparsity of $\widehat{\mathbf{\Omega}}$. The results are more sensitive to the choice of λ_1, since it directly influences the sparsity of the autoregressive parameters. It turns out that Sparse VAR still outperforms the other estimators for a large range of λ_1 values.

1.5 Data and Model

We use the sparse estimation technique for large VARs described in Section 1.3 to identify cross-category demand effects across 17 categories in the Dominick's Finer Foods database. This database is a well-established source of weekly scanner data from a large Midwestern supermarket chain, Dominick's Finer Foods (e.g. Kamkura and Kang, 2007, Pauwels, 2007). We first describe the data and model in more detail, and then report on the insights the Sparse VAR generates in the next section.

Table 1.2: *Description of the 17 categories from Dominick's Finer Foods database that are analyzed in this paper. For each category, we report the proportion of food and drink expenditures.*

Category	Expenditures	Category	Expenditures
Soft Drinks	22.24%	Snack Crackers	3.04%
Cereals	13.92%	Frozen Juices	2.88%
Cheeses	10.46%	Canned Tuna	2.80%
Refrigerated Juices	7.36%	Frozen Dinners	2.00%
Frozen Entrees	6.98%	Front-end-candies	2.00%
Beer	6.35%	Cigarettes	1.49%
Cookies	6.21%	Oatmeal	1.43%
Canned Soup	4.82%	Crackers	1.37%
Bottled Juices	4.66%		

We use all 17 product categories in the Dominick's Finer Foods database containing food and drink items, a much broader selection of categories than previous studies on cross-category demand effects have considered. A description of each product category can be found in Table 1.2. For 15 stores, we obtain weekly sales, pricing and promotional feature and display data for the 17 product categories.

Sales. Category sales volumes for the 17 categories, measured in dollars per week.

Promotion. The promotional data include the percentage of SKUs of each category that are promoted (feature and display) in a given week, following Srinivasan et al. [2004].

Prices. To aggregate pricing data from the SKU level to the product category level, we follow Srinivasan et al. [2004] and Pauwels et al. [2002] in using SKU market shares as weights. Prices are not deflated because there is strong evidence that people are sensitive to nominal rather than real price changes [Shafir et al., 1997] over short time periods.

We use data from January 1993 to July 1994, 77 weeks in total. We neither use data before 1993 since they contain missing observations, nor observations after 1994 since Srinivasan et al. [2004] pointed out that manufacturers made extensive use of 'pay-for-performance' price promotions as of 1994, which are not fully reflected in the Dominick's database. This data range is short relative to the dimension of the VAR, which calls for a regularization approach such as the Sparse VAR. For all stores, we collect data on sales, promotion and pricing for all

Table 1.3: *Description of the 15 data sets. Each data set contains multivariate time series for sales ($\boldsymbol{Y}_t$), promotion ($\boldsymbol{M}_t$) and prices ($\boldsymbol{P}_t$).*

Store	Number of	Dimension			
	Time Points	$\mathbf{Y}_t$	$\mathbf{M}_t$	$\mathbf{P}_t$	Total
Store 1-15	77	17	16	17	50

17 categories. Only for cigarettes, no promotion variable is included in the VAR since none of the SKUs in that category were promoted during the observation period.

We estimate a separate VAR model for each store, which allows to evaluate the robustness of the findings. The multivariate time series entering the VAR model are the log-differenced sales ($\mathbf{Y_t}$), differenced promotion ($\mathbf{M_t}$), and log-differenced prices ($\mathbf{P_t}$).[1] The dimensions of the time series are represented in Table 1.3. We use the Vector Autoregressive model, with endogenous promotion and prices,

$$\begin{bmatrix} \mathbf{Y}_t \\ \mathbf{M}_t \\ \mathbf{P}_t \end{bmatrix} = \mathbf{B}_0 + \mathbf{B}_1 \begin{bmatrix} \mathbf{Y}_{t-1} \\ \mathbf{M}_{t-1} \\ \mathbf{P}_{t-1} \end{bmatrix} + \ldots + \mathbf{B}_p \begin{bmatrix} \mathbf{Y}_{t-p} \\ \mathbf{M}_{t-p} \\ \mathbf{P}_{t-p} \end{bmatrix} + \mathbf{e}_t. \tag{1.6}$$

Averaged across stores, the selected value of p is two for the Sparse VAR. Also for the Bayesian estimators, the lag order of the VAR is selected using the BIC criterion, which is one for the majority of the stores.

1.6 Empirical Results

We focus on the effects of prices, promotions and sales in category A on the sales (or demand) in category B, where A and B belong to the product category network. We first study the direct effects. For instance, there is no direct effect of price of A on sales of B if the corresponding estimated regression coefficients are equal to zero at all lags. Then we turn to the complete chain of direct and indirect effects using Impulse Response Functions. For instance, price in category A indirectly influences sales in category B when the price of category A influences the price, promotion or sales in a certain other category C which, in turn, influences the sales of category B. Since we work in a time series setting, both direct

[1] Following standard practice, we first test for stationarity. A stationarity test of all individual time series using the Augmented Dickey-Fuller test indicates that most time series in levels are integrated of order 1.

Table 1.4: *Proportion of nonzero within and cross-category effects of price, promotion and sales on sales, averaged across 15 stores and 17 product categories.*

	Price	Promotion	Sales
Within-category	34%	30%	96%
Cross-category	19%	21%	21%

and indirect effects are dynamic in the sense that the effect occurs with a certain delay.

1.6.1 A network of product categories

We analyze cross-category demand effects as a network of interlinked product categories of which prices, promotions and sales in one category have an effect on sales in other categories. Recently, network perspectives have been increasingly used by marketing researchers to model, for example, the network value of a product in a product network [Oestreicher-Singer et al., 2013] or to investigate the flow of influence in a social network [Zubcsek and Sarvary, 2011]. In our case, the 17 product categories are the nodes of the network. We estimate the Sparse VAR for 15 stores separately. If the Sparse VAR estimation results indicate, by giving a non-zero estimate, that prices in one category have a direct influence on sales in another category in the majority of the 15 stores, a directed edge is drawn between them. The resulting directed network is plotted in Figure 1.1. Similarly, Figures 1.2 and 1.3 present cross-category effects of respectively promotion and sales on sales. If promotion or sales in one category directly influence sales in another category, respectively, this is indicated by a directed edge.

A first important finding is that the cross-category networks are sparse – not each category influences each and every other category. While the sparse VAR estimation favors zero-effects, it does not enforce them. Here, as many as 78% of all estimated effects are zero-effects. Table 1.4 summarizes the prevalence of within-and cross-category effects. As expected, within-category effects are more common than cross-category effects. For all categories, past values of the own category's sales are selected for almost all stores. Cross-category effects of price on sales (19%), promotion on sales (21%) and sales on sales (21%) are about equally prevalent.

Next, we focus on category influence and responsiveness in the cross-category network, measured by the number of edges originating from and pointing to a

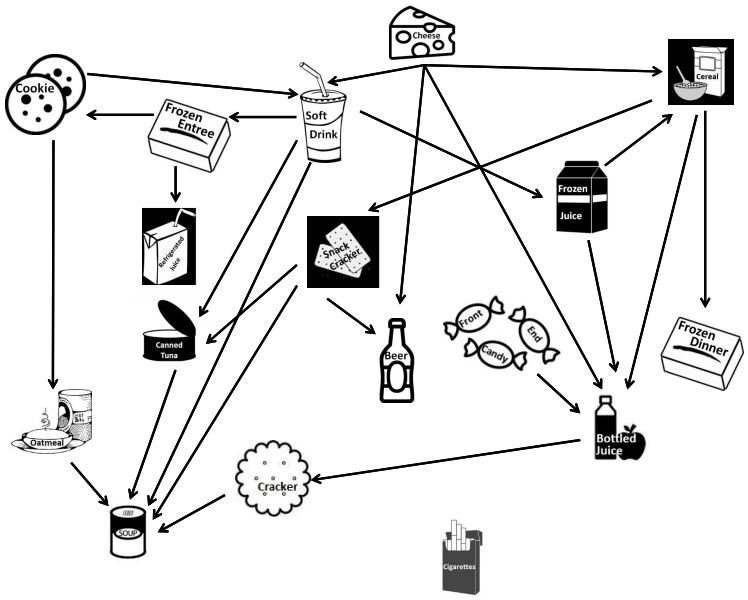


Figure 1.1: *Cross-category effect network of prices on sales: a directed edge is drawn from one category to another if its price influences sales in the other category for the majority of stores.*

category respectively. As discussed in Section 1.2, destination categories are expected to be more influential, while convenience categories are expected to be more responsive. We discuss which types of categories we find to be most influential and/or responsive in the cross-category networks of prices on sales, promotion on sales, and sales on sales.

The most influential categories in the cross-category network of prices on sales are destination categories such as Soft Drinks and Cheeses (cfr. each four outgoing edges in Figure 1.1). This is consistent with our expectations, as Soft Drinks is known to be a destination category [Briesch et al., 2013, Shankar and Kannan, 2014, Blattberg et al., 1995]. Soft Drinks is ranked first and Cheeses third in terms of food and drink expenditures (see Table 1.2) and are both heavily promoted by retailers. A price change in either of these categories thus strongly

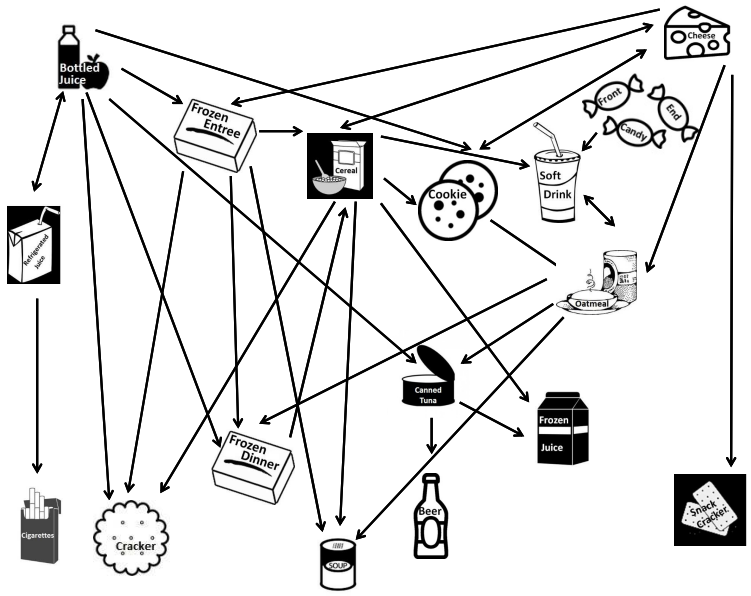


Figure 1.2: *Cross-category effect network of promotions on sales: a directed edge is drawn from one category to another if its promotion influences sales in the other category for the majority of stores.*

influences the budget constraint, which in turn influences purchase decisions in other categories. In the cross-category network of promotions on sales, Cereals is the most influential category (cfr. five outgoing edges in Figure 1.2). Briesch et al. [2013] identified Cereals as highly ranked among the destination categories. This is not surprising as cereals are part of daily consumption patterns and are ranked second in terms of food and drink expenditures. In the cross-category effects network of sales on sales in Figure 1.3, we identify again Cheeses as the most influential category.

We find convenience categories to be highly responsive to changes in other categories. The most prominent price effects are observed for Canned Soup (cfr. five incoming edges in Figure 1.1); the most prominent promotion effects for Frozen Dinners, Crackers and Canned Soup (cfr. each three incoming edges in

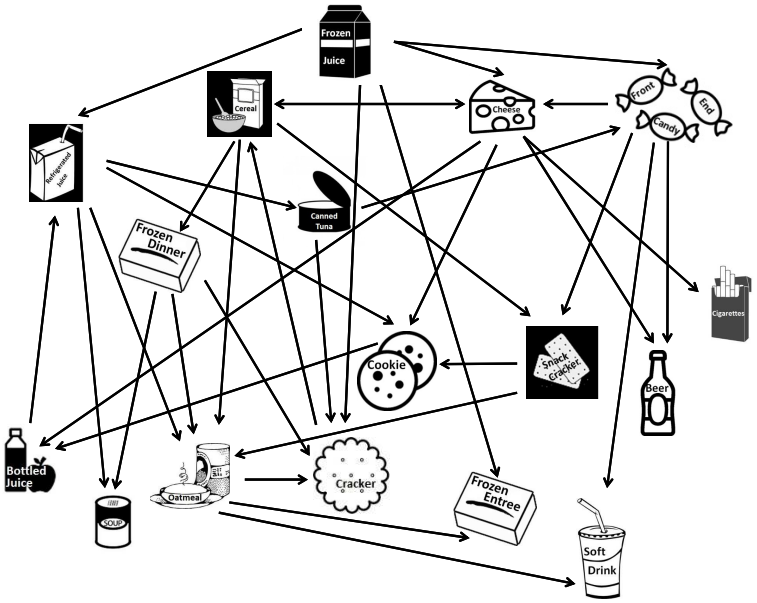


Figure 1.3: *Cross-category effect network of sales on sales: a directed edge is drawn from one category to another if its sales influences sales in the other category for the majority of stores.*

Figure 1.2); and the most prominent sales effects for Oatmeal and Crackers (cfr. each four incoming edges in Figure 1.3). These categories are typically bought out of convenience, such as Frozen Dinners and Canned Soup; or bought on occasion, such as Oatmeal and Crackers, counting for a very small percentage of food and drink expenditures (see Table 1.2).

Routine categories such as Bottled Juices, Refrigerated Juices, Frozen Juices and Cookies score moderate-to-high on category influence but are also responsive. This is in line with our expectation of many grocery categories being routine categories that are moderately influential and moderately responsive. Finally, the cigarettes category is least responsive and least influential. This finding is not surprising as cigarettes are addictive, hence, smokers probably have a stable consumption unrelated to food and drinks.

Table 1.5: *Kendall's coefficient of concordance across stores of cross-category effects of price, promotion and sales on sales for both category influence and responsiveness. P-values are indicated between parentheses.*

	Price	Promotion	Sales
Influence	0.40 (<0.001)	0.56 (<0.001)	0.30 (<0.001)
Responsiveness	0.30 (<0.001)	0.16 (0.001)	0.17 (<0.001)

To confirm the robustness of the results obtained by Sparse VAR, we check whether category responsiveness and influence are consistent across stores. We compute Kendall's coefficient of concordance W for category influence and responsiveness calculated from the graphs in Figures 2-4 at the store level. As W increases from 0 to 1, there is stronger consistency across stores. Table 1.5 indicates that all values of Kendall's W are significant.

1.6.2 Impulse Response Functions

For each store, we estimate the Sparse VAR and compute the corresponding Impulse Response Functions (IRFs). The effect size of an impulse is obtained by summing the absolute values of the responses across the first 10 lags of the IRF, where we take absolute values in order not to average out positive and negative response. We compute effect sizes of impulses in price, promotion or sales in one product category on the sales in the same (within) category or another (cross) category. In Table 1.6, we report the within and cross-category price, promotion and sales effect sizes, averaged across the 15 stores and the product categories.

Table 1.6 indicates that, for example, a one standard deviation price shock leads to an accumulated absolute change of .004 in own sales growth over a time period of 10 lags. As for the direct effects, we systematically find that within-category effects are larger in magnitude than cross-category effects, especially for sales and prices. For the marketing mix, promotions exert stronger within- as well as cross-category effects than price changes.

To get more insight in the sign of the cross-category effects, we summarize each IRF by the sum of the first 10 responses, and average this number over all stores. Table 1.7 reports the five largest positive and negative cross-category effects of price, promotion and sales on sales.

Table 1.6: *Size of within and cross-category effects of price, promotion and sales on sales, summed across 10 lags of the IRF, averaged across stores and product categories, and in absolute value.*

	Price	Promotion	Sales
Within-category	0.004	0.006	0.057
Cross-category	0.002	0.005	0.002

Cross-category price effects. We investigate whether consumers perceive categories as complements or as substitutes. Complementary and substitution effects occur between categories because they are consumed together or separately. Following the standard economic definition [Pashigian, 1998], complements are defined as goods having a negative cross-price elasticity, whereas substitutes are defined as goods having a positive cross-price elasticity. We find evidence of two important drivers of cross-category price effects: consumption relatedness and the budget constraint.

As an example of consumption relatedness, consider Soft Drink prices and Frozen Juices. An increase in Soft Drink prices makes consumers spend more on other drinks as a compensation, in particular Frozen Juices (see Table 1.7). The joint dynamic effect of a one standard deviation price impulse of Soft Drinks on the sales response growth of Frozen Juices is depicted in Figure 1.4 for the first three stores in the data set. Note that the instantaneous effect is estimated as exactly zero since the Sparse VAR puts the corresponding effect in the $\widehat{\mathbf{\Sigma}}$ matrix to zero. We see a sharp increase in Frozen Juices sales growth one week after the soft drink price increase, indicating substitution. However, the next two weeks, sales growth of Frozen Juices slows down, which could indicate stockpiling behavior [Gangwar et al., 2014].

Another example of consumption relatedness is Soft Drinks and Frozen Entrees. As can be seen from Table 1.7, we find a strong negative effect of Soft Drink prices on Frozen Entrees. This might be due to the fact that Soft Drinks and Frozen Entrees are consumed together. We do not find the opposite effect of price changes in Frozen Entrees on the sales of Soft Drinks. This asymmetry arises because Soft Drinks is a destination category (high influence), while Frozen Entrees is a convenience category (highly responsiveness).

Concerning the budget constraint, prominent cross-category price effects are observed for Soft Drinks and Cereals, both destination categories. Soft Drinks and Cereals account for a relatively large proportion of the expenditures of US

Table 1.7: *Cross-category price, promotion and sales effects on sales summed across 10 lags of IRFs and averaged across stores. We present only the five largest positive and negative effects.*

Cross-category price effects					
Price impulse	Sales response	Effect	Price impulse	Sales response	Effect
Perceived complements			Perceived substitutes		
Soft Drinks	Canned Tuna	-0.0209	Front-end-candies	Bottled Juices	0.0120
Soft Drinks	Frozen Entrees	-0.0182	Soft Drinks	Frozen Juices	0.0060
Canned Tuna	Canned Soup	-0.0173	Snack Crackers	Beer	0.0058
Cereals	Frozen Dinners	-0.0104	Cookies	Oatmeal	0.0056
Bottled Juices	Crackers	-0.0074	Frozen Juices	Bottled Juices	0.0023
Cross-category promotion effects					
Promotion impulse	Sales response	Effect	Promotion impulse	Sales response	Effect
Bottled Juices	Frozen Entrees	0.0586	Oatmeal	Canned Tuna	-0.0214
Cheeses	Frozen Entrees	0.0421	Cheeses	Cookies	-0.0160
Crackers	Frozen Entrees	0.0246	Bottles Juices	Canned Tuna	-0.0158
Frozen Dinners	Frozen Entrees	0.0170	Refrigerated Juices	Canned Tuna	-0.0128
Snack Crackers	Frozen Entrees	0.0127	Cereals	Cheeses	-0.0127
Cross-category sales effects					
Sales impulse	Sales response	Effect	Sales impulse	Sales response	Effect
Front-end-candies	Soft Drinks	0.0191	Snack Crackers	Oatmeal	-0.0154
Oatmeal	Frozen Entrees	0.0123	Frozen Juices	Frozen Entrees	-0.0120
Canned Tuna	Crackers	0.0094	Cereals	Frozen Dinners	-0.0099
Front-end-candies	Beer	0.0086	Snack Crackers	Cookies	-0.0087
Snack Crackers	Frozen Dinners	0.0064	Refrigerated Juices	Canned Tuna	-0.0084

families (respectively 22% and 14% of spending on food and drinks, see Table 1.2), which indicates that the budget constraint is an important source of cross-category effects.

Cross-category promotion effects. The results in Table 1.7 indicate that branding and promotion intensity are important drivers of cross-category promotion effects. Concerning branding, cross-category promotion effects are observed for categories that share brands such as Frozen Dinners and Frozen Entrees (e.g. the frozen prepared foods brand 'Stouffer's'). Concerning promotion intensity, prominent cross-category promotion effects are observed for categories in which a high percentage of the SKUs is promoted, such as Cheeses and Bottled Juices (respectively 28% and 26% of SKUs, on average, are promoted in our data.) A promotion impulse in such categories might either trigger joint consumption (e.g. Bottled Juices and Frozen Entrees), or deter consumption (e.g. Cheeses and Cookies).

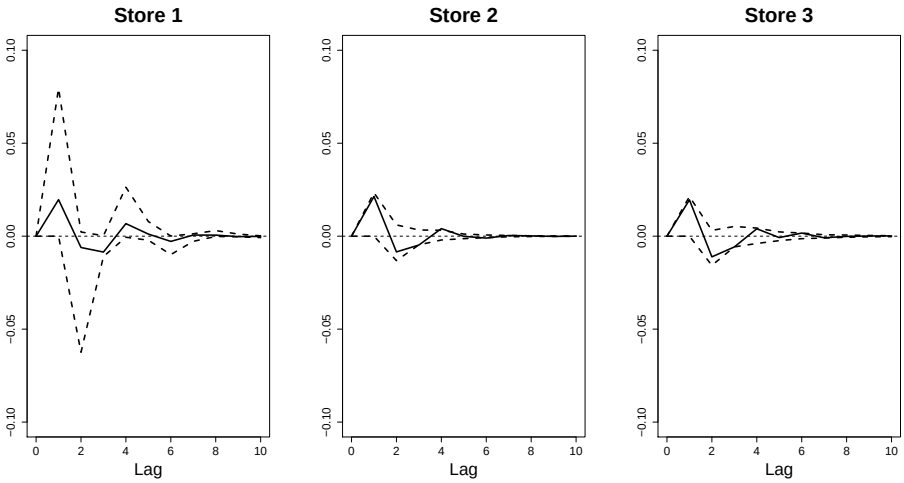


Figure 1.4: *Impulse response function: response of frozen juices sales growth to a one standard deviation impulse in the price of soft drinks.*

Cross-category sales effects. In Table 1.7, we find evidence of two important drivers of cross-category effects of sales on sales: affinity in consumption and the budget constraint. Prominent cross-category sales effects occur because of affinity in consumption. Some categories are jointly consumed towards a common goal, such as Front-end-candies and Soft Drinks/Beer (for a light meal); while others such as Snack Crackers and Cookies are purchased as replacements since consumers might perceive them to have a similar functionality. Concerning the budget constraint, we find some cross-category sales effects between seemingly unrelated categories such as Refrigerated Juices and Canned Tuna.

Importantly, the results from Table 1.7 are in line with our findings on category influence and responsiveness. Destination categories such as Soft Drinks, Cereals and Cheeses mainly influence sales in other categories through their price, promotion or sales impulses. Convenience categories such as Frozen Entrees and Frozen Dinners are more responsive to changes in other categories. Routine categories, such as Cookies, are moderately influential and moderately responsive, while occasional categories, such as Oatmeal, are highly responsive.

1.6.3 Robustness checks

Alternative penalty function. We investigate the robustness of the results to the choice of the penalty function. We re-estimate the models using the Sparse VAR with elastic net instead of the grouplasso penalty (a short explanation of the elastic net is given in Section 1.4). The managerial insights obtained by Sparse VAR with either grouplasso or elastic net are very similar. Similarities are that (i) within-category effects are more common and larger in magnitude than cross-category effects, (ii) destination categories such as Cheeses and Cereals are very influential, (iii) convenience categories such as Frozen Entrees, and occasional categories such as Crackers are very responsive (iv) routine categories such as Bottled Juices, Refrigerated Juices and Cookies are both influential and responsive (v) the most prominent cross-category effects of price, promotion and sales on sales are highly overlapping.

Alternative data period. We also check the performance of the Sparse VAR on the post-1994 data. Retailers made extensive use of 'pay-for-performance' price promotions that are not fully reflected in the Dominick's database. The data generating process might have changed in this period. Therefore, we should not assume constant parameter values. We re-estimate the model on the post-1994 data (data from October 1995 until May 1997) and verify its performance. In the post-1994 period, similar conclusions can be drawn with respect to within versus cross-category effects and category influence and responsiveness. Some differences are observed in the post-1994 period concerning the impulse response functions. These differences occur due to an altered strategy concerning average pricing and promotion intensity in the 17 product categories in the post-1994 period compared to the 1993-1994 period. Detailed results are available from the authors upon request.

Alternative price time series. We investigate the robustness of the results to the calculation of the price time series. Instead of aggregating prices from the SKU level to the product category level using SKU market shares as weights (cfr. Section 1.5), we now take the normal mean over all SKUs. We re-estimate the model using the newly calculated price time series. Similar insights are obtained with respect to cross-category effects and category influence and responsiveness. The most influential categories in the cross-category network of prices on sales are the destination categories Cereals and Cheeses; the most responsive ones the convenience categories Frozen Entrees, Frozen Dinners and Canned Tuna.

Alternative sparsity parameter selection. Our results are based on the BIC to

Table 1.8: *Mean Absolute Forecast Error (MAFE) for category-specific sales, averaged over the 15 stores and the 17 product categories. P-values of a Diebold-Mariano test comparing the Sparse VAR to its alternatives are indicated between parentheses.*

	Sparse VAR	LS	Restricted LS		Bayesian Methods	
			1-step	Iterative	Minnesota	NIW
MAFE	736.80	1298.54 (<0.01)	784.96 (<0.01)	734.82 (0.38)	875.47 (<0.01)	1078.03 (<0.01)

select the penalty parameters. We also ran the analysis using AIC as a selection criterion for the penalty function. While the model selected by AIC are slightly less sparse, the substantive insights do not change.

1.6.4 Forecast Performance

Although prediction is not the main goal of the proposed methodology, we deem it important to show that the Sparse VAR can compete with other methods in terms of prediction accuracy. We estimate model (1.6) for each store and perform a forecast exercise (cfr. Section 1.4), using a rolling window of length $S = 67$. One-step-ahead forecasts of sales for each product category are computed for $t = S, \ldots, T-1$, with $T = 77$. The same estimation methods as in Section 1.4 are used.

Results on the sales predictions are summarized in Table 1.8 by the Mean Absolute Forecast Error (MAFE), averaged across time and over the 17 product categories and 15 stores. The MAFE should be seen as a measure of forecast accuracy, not as a measure of managerial relevance of the obtained results. The variable selection methods Sparse VAR, 1-step and Iterative Restricted LS perform, on average, better than the methods that do not perform variable selection. This indicates that sparsity improves prediction accuracy. Sparse VAR and Iterative Restricted LS achieve the best forecasting performance. A Diebold-Mariano test confirms that latter two methods significantly outperform the other methods. We conclude that the improvement in interpretability of the model obtained by Sparse VAR, as discussed in the previous section, does not come at the cost of lower forecast performance.

1.7 Discussion

This paper presents a Sparse VAR methodology to detect the inter-relationships in a large product category network. In the cross-category demand effects application, we detect an important number of cross-category demand effects for a large number of categories. We find that categories have asymmetric roles: While destination categories are more influential, convenience categories are more responsive. We identify main perceived cross-category effects but also detect cross-category effects between categories that are not directly related at first sight. Hence, the need to study – potentially a large number of – product categories simultaneously. While cross-category effects are prevalent, many of them are still absent, calling for a sparse estimation procedure that succeeds in highlighting the main inter-relationships in the product category network.

Our finding on the asymmetric roles of categories in the product category networks is in line with the analysis of Bonfrer et al. [2006]. While economic theory implies compensated price effects to be pairwise symmetric with respect to their magnitude (e.g. Deaton and Muellbauer, 1980), it does not imply magnitude symmetry with respect to compensated cross-price *elasticities*. Asymmetries in cross-price elasticities might be explained by differences in the budget share-weighted income elasticity and/or the category demand elasticity of two categories [Bonfrer et al., 2006].

We identify category influence and responsiveness in our cross-category demand effects application using aggregate store level data. Other cross-category studies, such as Russell and Kamakura [1997], Ainslie and Rossi [1998], Russell et al. [1999], Russell and Petersen [2000], Elrod et al. [2002] use market basket data. Since the availability and use of such market basket data pose difficulties to managers, they rarely use market basket data for category analysis [Shankar and Kannan, 2014]. As managerial decisions are often made at the category level, managers prefer to work with more readily available aggregate store level data. Hence, using aggregate category store level is managerially relevant [Ailawadi et al., 2009, Leeflang and Selva, 2012].

A first limitation of our approach is that we use aggregate category data, which might lead to biased estimates when there is heterogeneity on the SKU level [Dekimpe and Hanssens, 2000]. Second, our model does not allow to estimate cross-category effects on the individual consumer level. Insights into the behavior of consumers are revealed using market basket data, which requires a very different modeling approach. Despite these limitations, aggregate category data are highly

relevant from the perspective of category management within the store.

An important advantage of the Sparse VAR is that it overcomes the dimensionality problem – it results in a parsimonious model with minimal structural constraints. We show that this leads to more accurate estimation and prediction results as compared to standard Least Squares methods. If the researcher wishes to restrict some of the parameters to zero a priori, using marketing theory, this is of course still possible to implement with the Sparse VAR. The same holds for the reverse, i.e. forcing some variables to be included in the model, which can be done by adjusting the penalty on the regression coefficients in (1.3).

The methodology presented in this paper is relevant in a variety of other settings. First, Sparse VAR can be used to study competitive demand effects across many competitors. The VAR is ideal for measuring competitive effects since it is able to capture own- and cross-elasticity of sales to both pricing and marketing spending [Srinivasan et al., 2004, Horvath et al., 2005]. Typically only three competitors are included in such studies, while using the Sparse VAR allows for a much larger number to be included. Second, in the field of international marketing research there is an increased interest in studying cross-country spill-over effects, as for example in Albuquerque et al. [2007], van Everdingen et al. [2009] and Kumar and Krishnan [2002]. Every country that is added to the data set leads to an increase in the number of cross-country parameters to be estimated. Using the proposed methodology, a large VAR model could be built which allows spill-over effects between many countries. Finally, the Market Response Model could be extended with data on online word of mouth or online search, which are now readily available. Especially in the Big Data era, most companies collect an abundance of variables [Chintagunta et al., 2013], such that large VAR models will become even larger as more granular data become available.

1.8 Appendix: Penalized Likelihood Estimation

We iteratively solve the minimization problem (1.3) for $\boldsymbol{\beta}$ conditional on $\boldsymbol{\Omega}$ and then for $\boldsymbol{\Omega}$ conditional on $\boldsymbol{\beta}$.

Solving for $\boldsymbol{\beta}|\boldsymbol{\Omega}$: When $\boldsymbol{\Omega}$ is fixed, the minimization problem in (1.3) is equivalent to minimizing

$$\widehat{\boldsymbol{\beta}}|\boldsymbol{\Omega} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^T(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) + \lambda_1 \sum_{g=1}^{G} ||\boldsymbol{\beta}_g||_2 \,, \tag{1.7}$$

where $\widetilde{\mathbf{y}} = \mathbf{P}\mathbf{y}$, $\widetilde{\mathbf{X}} = \mathbf{P}\mathbf{X}$, and $\mathbf{P}$ is a matrix such that $\mathbf{P}^T\mathbf{P} = \widetilde{\mathbf{\Omega}}$. The transformation of the data to $\widetilde{\mathbf{y}}$ and $\widetilde{\mathbf{X}}$ ensures that the resulting model has uncorrelated and homoscedastic error terms. The above minimization problem is convex if $\mathbf{\Omega}$ is nonnegative definite. The minimization problem is equivalent to the groupwise lasso of Yuan and Lin [2006], implemented in the R package `grplasso` [Meier, 2009].

Solving for $\mathbf{\Omega}|\boldsymbol{\beta}$: When $\boldsymbol{\beta}$ is fixed, the minimization problem in (1.3) reduces to

$$\widehat{\mathbf{\Omega}}|\boldsymbol{\beta} = \underset{\mathbf{\Omega}}{\operatorname{argmin}} \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\widetilde{\mathbf{\Omega}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \log|\mathbf{\Omega}| + \lambda_2 \sum_{k \neq k'} |\Omega_{kk'}|\,, \tag{1.8}$$

which corresponds to penalized covariance estimation. Using the glasso algorithm of Friedman et al. [2008], available in the R package `glasso` [Friedman et al., 2011], the optimization problem in (1.8) is solved.

We start the algorithm by taking $\widehat{\mathbf{\Omega}} = \mathbf{I}_q$ and iterate until convergence. We iterate until $\max_s|\hat{\beta}_{s,i} - \hat{\beta}_{s,i-1}| < \epsilon$, with $\hat{\beta}_{s,i}$ the s^{th} parameter estimate in iteration i (same for $\hat{\mathbf{\Omega}}$) and the tolerance ϵ set to 10^{-3}.

Selecting the Sparsity Parameters and the order of the VAR. We first determine the optimal values of λ_1 and λ_2 for a fixed value of p, the order of the VAR. The sparsity parameters λ_1 and λ_2 are selected according to a minimal Bayes Information Criterion (BIC). In the iteration step where $\boldsymbol{\beta}$ is estimated conditional on $\mathbf{\Omega}$, we solve (1.7) over a range of values for λ_1 and select the one with lowest value of

$$BIC_{\lambda_1} = -2\log L_{\lambda_1} + k_{\lambda_1}\log(n), \tag{1.9}$$

where L_{λ_1} is the estimated likelihood, corresponding to the first term in (1.7), using sparsity parameter λ_1. Furthermore, k_{λ_1} is the number of non-zero estimated regression coefficients and n the number of observations. Similarly, for selecting λ_2, we use the BIC given by

$$BIC_{\lambda_2} = -2\log L_{\lambda_2} + k_{\lambda_2}\log(n)\,. \tag{1.10}$$

Finally, we select the order p of the VAR. We estimate the VAR for different values of p. The optimal values of λ_1 and λ_2 are determined for a each of those values of p. We select the order p of the VAR using BIC:

$$BIC_{(p,\lambda_1(p),\lambda_2(p))} = -2\log L_{(p,\lambda_1(p),\lambda_2(p))} + k_{(p,\lambda_1(p),\lambda_2(p))}\log(n)\,, \tag{1.11}$$

where $L_{(p,\lambda_1(p),\lambda_2(p))}$ and $k_{(p,\lambda_1(p),\lambda_2(p))}$ depend on the value p and the optimally chosen values of $\lambda_1(p)$ and $\lambda_2(p)$ for that specific value of p.

Chapter 2

An algorithm for the multivariate grouplasso with covariance estimation

Abstract

We study a grouplasso estimator for the multivariate linear regression model that accounts for correlated error terms. A block coordinate descent algorithm is used to compute this estimator. We perform a simulation study with categorical data and multivariate time series data, typical settings with a natural grouping among the predictor variables. Our simulation studies show the good performance of the proposed grouplasso estimator compared to alternative estimators. We illustrate the method on a time series data set of gene expressions.

2.1 Introduction

Since its introduction by Yuan and Lin [2006], the group least absolute shrinkage and selection operator (grouplasso) has received considerable interest in the statistical literature (e.g. Meier et al., 2008, Wang and Leng, 2008, Peng et al., 2010, Simon et al., 2013, Alfons et al., 2016). In many applications, the parameter vector in the regression model is structured into groups. Typical examples are (i) regression with categorical variables, where a group of dummies represents each categorical variable, or (ii) time series regression where several lagged values of

the same time series are included in the model. In settings with such a natural group structure, one wants to select either all or none of the variables belonging to a particular group. The key strength of the grouplasso lies in its ability to perform such groupwise selection.

We consider the grouplasso for the multivariate linear regression model. The multivariate linear regression model generalizes the classical linear regression model in that it regresses $q > 1$ responses instead of a single response on p predictors. Let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_q) \in \mathbb{R}^{n \times q}$ be the response matrix, and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$ be the predictor matrix. The error vectors are assumed to follow a normal $N_q(0, \mathbf{\Sigma})$ distribution, with $\mathbf{\Sigma}^{-1} = \mathbf{\Omega}$, and are collected in the columns of the error matrix $\mathbf{E}$. The multivariate linear regression model is given by

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{2.1}$$

where $\mathbf{B} \in \mathbb{R}^{p \times q}$ is the coefficient matrix. We assume that this coefficient matrix contains K predefined groups. Denote each group G_j containing one or more elements of $\mathbf{B}$ as $\mathbf{B}_{G_j}$ where $j \in \{1, \ldots, K\}$.

Recently, Li et al. [2015] discussed the grouplasso for the multivariate linear regression model. Their multivariate grouplasso estimator[1] is given by

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}}{\operatorname{argmin}} \frac{1}{2n} \operatorname{tr}\left((\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\right) + \sum_{j=1}^{K} \lambda_{G_j} m_j ||\mathbf{B}_{G_j}||_2, \tag{2.2}$$

where $\operatorname{tr}(\cdot)$ denotes the trace, $\lambda_{G_j} > 0$, for $1 \leq j \leq K$, are sparsity parameters, and m_j equals the number of elements in group j. A groupwise penalty is used for the regression coefficients. As such, variables are selected in a grouped manner: either all elements of a certain group are set to zero or none.

However, Li et al. [2015] do not account for correlated errors. Accounting for correlated errors has been found to increase estimation accuracy, see e.g. Rothman et al. [2010] for the multivariate lasso with covariance estimation or Chen and Huang [2016] for sparse multivariate reduced rank regression with covariance estimation. We therefore extend the multivariate grouplasso from Li et al. [2015] such that the correlation between the error terms of the different equations of the multivariate regression model is taken into account. To this end, we simultaneously estimate the regression parameters $\mathbf{B}$ and the inverse covariance matrix of

[1] Note that Li et al. [2015] consider a more general version of the multivariate grouplasso that also allows for selection of predictors within the important groups.

the error terms $\mathbf{\Omega}$ using penalized maximum likelihood:

$$(\widehat{\mathbf{B}}, \widehat{\mathbf{\Omega}}) = \underset{(\mathbf{B},\mathbf{\Omega})}{\operatorname{argmin}} \frac{1}{2n} \operatorname{tr}\left((\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\mathbf{\Omega}\right) - \frac{1}{2}\log|\mathbf{\Omega}| + \\ + \sum_{j=1}^{K} \lambda_{G_j} m_j ||\mathbf{B}_{G_j}||_2 + \lambda_\omega \sum_{k \neq k'} |\omega_{kk'}|, \quad (2.3)$$

where $\lambda_\omega > 0$ is a sparsity parameter, and ω_{kk} is the k^{th} element of $\mathbf{\Omega}$. We use an L_1 penalty for the elements of the inverse covariance matrix.

Section 2.2 describes the algorithm used to approximate the minimizer of the objective function in (2.3). The main modification in the algorithm compared to the proposal of Li et al. [2015] is that the error covariance structure is taken into account. Simulation studies are performed in Section 2.3. Our simulations show that the grouplasso with covariance estimation considerably outperforms the grouplasso without covariance estimation. Section 2.4 contains a real data example.

2.2 The algorithm

To find the minimum of the penalized negative log-likelihood in (2.3), we iteratively solve for $\mathbf{B}$ conditional on $\mathbf{\Omega}$ and for $\mathbf{\Omega}$ conditional on $\mathbf{B}$.

Solving for $\mathbf{B}$ *conditional on* $\mathbf{\Omega}$. When $\mathbf{\Omega}$ is fixed, the minimization problem in (2.3) is equivalent to

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}}{\operatorname{argmin}} \frac{1}{2n} \operatorname{tr}\left((\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\mathbf{\Omega}\right) + \sum_{j=1}^{K} \lambda_{G_j} m_j ||\mathbf{B}_{G_j}||_2. \quad (2.4)$$

To find a solution to (2.4), we use a block coordinate descent algorithm, analogously to Friedman et al. [2008] for solving the single response lasso problem, or to Li et al. [2015] for the multivariate grouplasso problem without covariance estimation. Lemma 1 (Lemma 4.2 from Chapter 4 in Bühlmann and van de Geer, 2011) provides a necessary and sufficient condition for $\widehat{\mathbf{B}}$ to be a solution of (2.4).

Lemma 1. *Denote the loss function by*

$$\rho(\mathbf{B}) = \frac{1}{2n} \operatorname{tr}\left((\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\mathbf{\Omega}\right).$$

The gradient of the loss function evaluated at $\mathbf{B}$ *is*

$$\nabla \boldsymbol{\rho}(\mathbf{B}) = -\frac{\mathbf{1}}{\mathbf{n}} \mathbf{X^T}(\mathbf{Y} - \mathbf{XB})\mathbf{\Omega}.$$

A necessary and sufficient condition for $\mathbf{B}$ *to be a solution of* (2.4) *is*

(i) $\nabla \boldsymbol{\rho}(\mathbf{B})_{G_j} + \lambda_{G_j} m_j \dfrac{\mathbf{B}_{G_j}}{||\mathbf{B}_{G_j}||_2} = \mathbf{0}$ *if* $\mathbf{B}_{G_i} \neq \mathbf{0}$

(ii) $||\nabla \boldsymbol{\rho}(\mathbf{B})_{G_j}||_2 \leq \lambda_{G_j} m_j$ *if* $\mathbf{B}_{G_j} = \mathbf{0}$.

To start up the block coordinate descent algorithm, an initial value for $\mathbf{B}$ is needed. We use the lasso estimator obtained by performing q separate lasso regressions.[2] Assume now that $\widehat{\mathbf{B}}^{(m-1)}$ is given, for $m \geq 1$. In the following iteration step m, we update our estimate from $\widehat{\mathbf{B}}^{(m-1)}$ to $\widehat{\mathbf{B}}^{(m)}$. Note that the ik^{th} element of the gradient of the loss function evaluated at $\mathbf{B}$ is given by

$$\begin{aligned}
\nabla \boldsymbol{\rho}(\mathbf{B})_{ik} &= -\frac{1}{n}\mathbf{x}_i^T(\mathbf{Y} - \mathbf{X}\mathbf{B})\boldsymbol{\Omega}_k \\
&= \frac{1}{n}\left(-\mathbf{x}_i^T(\mathbf{Y} - \mathbf{X}\mathbf{B}^{-ik})\boldsymbol{\Omega}_k + \omega_{kk}||\mathbf{x}_i||_2^2 B_{ik}\right) \\
&= \frac{1}{n}\left(-\mathbf{S}_{ik} + \omega_{kk}||\mathbf{x}_i||_2^2 B_{ik}\right),
\end{aligned}$$

with $\mathbf{x}_i$ the i^{th} column of $\mathbf{X}$, $\boldsymbol{\Omega}_k$ the k^{th} row of $\boldsymbol{\Omega}$, ω_{kk} the kk^{th} element of $\boldsymbol{\Omega}$, $\mathbf{B}^{-ik}$ is $\mathbf{B}$ with element ik replaced by zero, and $\mathbf{S}_{ik} = \mathbf{x}_i^T(\mathbf{Y} - \mathbf{X}\mathbf{B}^{-ik})\boldsymbol{\Omega}_k$.

In iteration step m, we cycle through all groups G_j, with $j = 1, \ldots, K$. If, for group G_j

$$||\nabla \boldsymbol{\rho}(\widehat{\mathbf{B}}^{(m-1)})_{G_j}||_2 \leq \lambda_{G_j} m_j$$

holds, then according to condition 2 from Lemma 1, all elements of group G_j of $\widehat{\mathbf{B}}^{(m)}$ are set to zero. Otherwise, according to condition 1 from Lemma 1, for every element ik of $\mathbf{B}$ belonging to group G_j it needs to hold that

$$\begin{aligned}
0 &= \nabla \boldsymbol{\rho}(\mathbf{B})_{ik} + \lambda_{G_j} m_j \frac{B_{ik}}{||\mathbf{B}_{G_j}||_2} \\
\iff \quad 0 &= \frac{-\mathbf{S}_{ik}}{n} + \frac{\omega_{kk}||\mathbf{x}_i||_2^2}{n} B_{ik} + \frac{\lambda_{G_j} m_j}{||\mathbf{B}_{G_j}||_2} B_{ik} \\
\iff B_{ik} &= \frac{\mathbf{S}_{ik}}{\omega_{kk}||\mathbf{x}_i||_2^2 + \dfrac{n\lambda_{G_j} m_j}{||\mathbf{B}_{G_j}||_2}}. \qquad (2.5)
\end{aligned}$$

The right-hand-side from equation (2.5) involves B_{ik} in the computation of $||\mathbf{B}_{G_j}||_2$. For this, we use the estimate from the previous iteration. Table 2.1 provides a schematic overview of the block coordinate descent algorithm.

[2] If the initial lasso estimator puts all elements of group j to zero - which occurs with a small probability - then all elements from group j will remain zero in the remainder of the algorithm. To limit this influence of the initial estimator, one could consider using the ridge estimator [Hoerl and Kennard, 1970] as initial estimator.

Table 2.1: *Block Coordinate Descent Algorithm to solve for* $\mathbf{B}$ *conditional on* $\mathbf{\Omega}$.

1: **Initialization** Let $\mathbf{B}^{(0)}$ be an initial parameter estimate. We use the lasso estimator obtained by performing q separate lasso regressions. Set $m = 0$.

2: **Repeat**

$m \leftarrow m + 1$

For each block $j = 1, \ldots, K$:

If $||\nabla \boldsymbol{\rho}(\widehat{\mathbf{B}}^{(m-1)})_{G_j}||_2 \leq \lambda_{G_j} m_j$: set $\widehat{\mathbf{B}}^{(m)}_{G_j} = \mathbf{0}$

Else: Update every ik^{th} element $\widehat{B}^{(m)}_{ik}$ of $\widehat{\mathbf{B}}^{(m)}$ belonging to group G_j by

$$\widehat{B}^{(m)}_{ik} = \frac{\widehat{S}^{(m-1)}_{ik}}{\omega_{kk}||\mathbf{x}_i||_2^2 + \dfrac{n\lambda_{G_j} m_j}{||\widehat{\mathbf{B}}^{(m-1)}_{G_j}||_2}}.$$

3: **Until** convergence. We iterate until the relative change in the value of the objective function in (2.4) in two successive iterations is smaller than the tolerance value $\epsilon = 10^{-2}$.

Note that the estimator in (2.4) is a multivariate *adaptive* grouplasso estimator since each group has its own sparsity parameter λ_{G_j}. We take $\lambda_{G_j} = \lambda/||\widehat{\mathbf{B}}^{(0)}_{G_j}||_2$, for $j = 1, \ldots, K$. This way, only one tuning parameter for the regression coefficients needs to be selected instead of K. We use a grid of sparsity parameters and search for the optimal one using the Bayesian Information Criterion (BIC). The BIC is given by

$$BIC_\lambda = -2\,\text{logL}_\lambda + k_\lambda \log(n),$$

where logL_λ is the estimated log-likelihood, corresponding to the first term of the objective function in (2.4), using sparsity parameter λ, and k_λ is the number of non-zero estimated regression coefficients.

Solving for $\mathbf{\Omega}$ *conditional on* $\mathbf{B}$. When $\mathbf{B}$ is fixed, the minimization problem in (2.3) corresponds to the graphical lasso [Friedman et al., 2008] on the residuals $\mathbf{Y} - \mathbf{XB}$. We use the Bayesian Information Criterion to select the optimal value of the sparsity parameter λ_ω (e.g. Yuan et al., 2007).

Starting value and convergence. We start by taking $\mathbf{\Omega} = \mathbf{I}$ and then iteratively solve for $\mathbf{B}$ conditional on $\mathbf{\Omega}$ and for $\mathbf{\Omega}$ conditional on $\mathbf{B}$. We iterate until the relative change in the value of the objective function in (2.3) in two successive iterations is smaller than the tolerance value $\epsilon = 10^{-2}$.

2.3 Simulation

We compare the performance of the multivariate grouplasso with covariance estimation, '`GroupLasso+Cov`', to

(i) The multivariate grouplasso without covariance estimation, '`GroupLasso`', i.e. the solution of (2.2),

(ii) The multivariate lasso with covariance estimation, '`Lasso+Cov`', i.e. the solution of (2.3) with $m_j = 1$, for $1 \leq j \leq K$, where $K = p \times q$. The resulting estimator is equivalent to the Multivariate Lasso With Covariance Estimator introduced in Rothman et al. [2010].

(iii) The multivariate lasso without covariance estimation, '`Lasso`', i.e. the solution of (2.2) with $m_j = 1$, for $1 \leq j \leq K$, where $K = p \times q$.

Note that '`Lasso+Cov`' and '`Lasso`' do not take the group structure among the predictors into account.

2.3.1 Predictor groups

The first data configuration corresponds to a regression model with categorical predictors, the second to a time series model.

Categorical data. We consider a design similar to model I from Yuan and Lin [2006] for the univariate regression model. We generate a sample Z_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, K$, of size n from a centered multivariate normal distribution with covariance matrix $\mathbf{\Sigma}^Z$ where

$$\Sigma^Z_{ij} = 0.5^{|i-j|}.$$

Afterwards, Z_{ij} is trichotomized as

$$C_{ij} = \begin{cases} 0 & \text{if} \quad Z_{ij} < \Phi^{-1}(\frac{1}{3}) \\ 1 & \text{if} \quad Z_{ij} > \Phi^{-1}(\frac{2}{3}) \\ 2 & \text{if} \quad \Phi^{-1}(\frac{1}{3}) < Z_{ij} < \Phi^{-1}(\frac{2}{3}), \end{cases}$$

for $i = 1, \ldots, n = 50$ and $j = 1, \ldots, K$, where K denotes the number of groups. We take $K \in \{5, 20, 50\}$. The $(n \times p)$ matrix of predictors $\mathbf{X}$ then contains in its columns the $p = 2K$ dummy variables $D^0_{ij} = \mathrm{I}(C_{ij} = 0)$ and $D^1_{ij} = \mathrm{I}(C_{ij} = 1)$,

for $j = 1, \ldots, K$ and $i = 1, \ldots, n$, where $\mathrm{I}(\cdot)$ is the indicator function. Next, the $q = 5$ responses are simulated from

$$\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{E}, \tag{2.6}$$

where $\mathbf{B} = \mathbf{I}_q \otimes \mathbf{b}$, with $\mathbf{b} = (2, -1, \ldots, 2, -1)$ a vector of length p/q. For the error covariance matrix $\mathbf{\Sigma}$ we consider different structures, detailed in the Section 2.3.2. The grouplasso accounts for the grouped predictor variables by selecting either all or none of the dummy variables corresponding to a particular categorical variable in one of the equations of the multivariate regression model.

Time series. We generate the data from a VAR(2) model

$$\mathbf{y}_t = \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_2\mathbf{y}_{t-2} + \mathbf{e}_t, \tag{2.7}$$

for $t = 1, \ldots, T = 50$, where $\mathbf{y}_t$ is a q-dimensional vector, with $q \in \{5, 20, 50\}$. The coefficient matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ have the same sparse structure and $\mathbf{e}_t \sim N_q(\mathbf{0}, \mathbf{\Sigma})$. For the error covariance matrix $\mathbf{\Sigma}$ we consider different structures, detailed in Section 2.3.2.

The above model is a Vector AutoRegressive (VAR) model of order two since two lagged values of each time series are included as predictors. The grouplasso accounts for the grouped predictor variables by selecting either all or none of the lagged values of a particular time series in one of the equations of the VAR. As a result, $\widehat{\mathbf{B}}_1$ and $\widehat{\mathbf{B}}_2$ have their zero elements in exactly the same cells.

We generate the sparse coefficient matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ from a network structure (see Fujita et al., 2007). This dimensions of this network are similar to the ones in the real data example to be discussed in Section 2.4. The adjacency matrix $\mathbf{A}$ represents the network structure where the nodes are the q different time series. Element $A_{ij} = 1$ if a directed edge is drawn from node i to node j, otherwise $A_{ij} = 0$. To construct the adjacency matrix $\mathbf{A}$, we start (iteration $l = 0$) from a network of two randomly selected nodes that are connected with a bidirectional edge. Next, in iteration $l = 1, \ldots, q-2$, a node that is currently not in the network is randomly selected. This new node is connected to a node that is present in the network via an edge whose direction is randomly chosen. The probability

$$\pi_m^{(l-1)} = \frac{d_m^{(l-1)}}{\sum_n d_n^{(l-1)}},$$

that the new node is connected to node m depends on the degree $d_m^{(l-1)}$ of the node present in the network from iteration $l-1$. The degree of a node equals the number of edges starting from it. Finally, we set $\mathbf{B}_1 = 0.4\mathbf{A}$ and $\mathbf{B}_2 = 0.2\mathbf{A}$.

2.3.2 Structure of the error terms

We consider three structures for the error covariance matrix $\mathbf{\Sigma}$ and its inverse $\mathbf{\Omega}$, see e.g. Rothman et al. [2010]:

(i) **Sparse $\mathbf{\Omega}$**: $\Sigma_{ij} = \rho^{|i-j|}$, with $\rho \in \{0.2, 0.4, 0.6, 0.8\}$. The error covariance matrix $\mathbf{\Sigma}$ is a dense matrix, whereas its inverse $\mathbf{\Omega}$ is a band matrix.

(ii) **Diagonal $\mathbf{\Omega}$**: $\mathbf{\Sigma} = \mathbf{I}_q$. Both the error covariance matrix and its inverse are diagonal.

(iii) **Dense $\mathbf{\Omega}$**: $\Sigma_{ij} = 0.5((|i-j|+1)^{2\times 0.9} - 2|i-j|^{2\times 0.9} + (|i-j|-1)^{2\times 0.9})$. Both the error covariance matrix and its inverse have a dense structure.

2.3.3 Performance measures

We measure estimation accuracy by looking at the Mean Absolute Estimation Error given by

$$\text{MAEE} = \frac{1}{N}\frac{1}{p\times q}\sum_{m=1}^{N}\sum_{j=1}^{q}\sum_{i=1}^{p}|\widehat{b}_{ij}^{(m)} - b_{ij}|, \tag{2.8}$$

where $\widehat{b}_{ij}^{(m)}$ is the estimate of the ij^{th} element of $\mathbf{B}$ in simulation run m. We take $N = 1000$ simulation runs.

We measure sparsity recognition by looking at the True Positive Rate and the True Negative Rate given by

$$\text{TPR} = \frac{1}{N}\sum_{m=1}^{N}\frac{\#\{(i,j) : \hat{b}_{ij}^{(m)} \neq 0 \ and \ b_{ij} \neq 0\}}{\#\{(i,j,) : \ b_{ij} \neq 0\}}$$
$$\text{TNR} = \frac{1}{N}\sum_{m=1}^{N}\frac{\#\{(i,j) : \hat{b}_{ij}^{(m)} = 0 \ and \ b_{ij} = 0\}}{\#\{(i,j,) : \ b_{ij} = 0\}}.$$

TPR gives the hit rate of including an important variable, whereas the TNR gives the hit rate of excluding an unimportant variable. Both should be as large as possible for reliable variable selection.

2.3.4 Results

In this section, we discuss the results for the two data configurations. We show that the `GroupLasso+Cov` considerably improves the `GroupLasso` as soon as the errors are correlated.

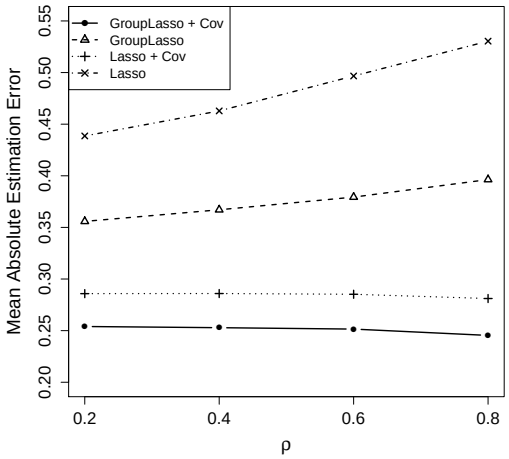


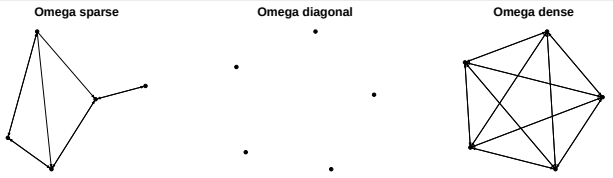
Figure 2.1: *Multivariate regression with $q = 5$ responses, $K = 5$ categorical regressors and $n = 50$: Mean Absolute Estimation Error versus the correlation ρ, for the four considered estimators.*

Categorical data. We first discuss the results for the sparse inverse error covariance structure (cfr. Section 2.3.2). The MAEE with $K = 5$ categorical regressors is displayed in Figure 2.1 for different values of the correlation ρ. Similar conclusion can be made for $K = 20$ or $K = 50$ categorical regressors and are, hence, omitted.

The `GroupLasso+Cov` substantially outperforms the `GroupLasso` for all values of the correlation ρ. The margin by which the former outperforms the latter increases when ρ increases. The `GroupLasso+Cov` achieves this improved estimation accuracy since it accounts for the error correlation whereas the `GroupLasso` does not. Besides, as expected for grouped predictors, the grouplasso estimators outperform the corresponding lasso estimators.

The MAEE for all simulation designs are reported in Table 2.2. In line with Figure 2.1, `GroupLasso+Cov` provides a considerable improvement in MAEE over `GroupLasso` when the error terms are correlated, see 'Omega sparse', with $\rho = 0.6$. For reasons of brevity, we only report the results for $\rho = 0.6$. The estimation accuracy improves by more than 30%. The improvement of `GroupLasso+Cov` over `GroupLasso` becomes even larger when the number of categorical regressors K increases. A paired t-test confirms that this improvement is significant (all

Table 2.2: *Multivariate regression with $q = 5$ responses, $K \in \{5, 20, 50\}$ categorical regressors and $n = 50$: Mean Absolute Estimation Error, True Positive and True Negative Rate.*



		$\rho = 0.6$					
	Estimator	MAEE	TPR/TNR	MAEE	TPR/TNR	MAEE	TPR/TNR
$K = 5$	`GroupLasso+Cov`	0.251	1.00/0.62	0.253	1.00/0.61	0.244	1.00/0.67
	`GroupLasso`	0.379	1.00/0.53	0.349	1.00/0.54	0.394	1.00/0.53
	`Lasso+Cov`	0.285	0.91/0.90	0.286	0.91/0.88	0.282	0.91/0.91
	`Lasso`	0.497	0.96/0.67	0.303	0.91/0.86	0.374	0.91/0.85
$K = 20$	`GroupLasso+Cov`	0.156	1.00/0.35	0.155	1.00/0.35	0.155	1.00/0.35
	`GroupLasso`	0.470	1.00/0.15	0.411	1.00/0.36	0.503	1.00/0.15
	`Lasso+Cov`	0.281	0.96/0.69	0.279	0.96/0.69	0.281	0.96/0.69
	`Lasso`	0.547	0.96/0.56	0.436	0.96/0.57	0.589	0.96/0.56
$K = 50$	`GroupLasso+Cov`	0.196	1.00/0.40	0.196	1.00/0.40	0.196	1.00/0.40
	`GroupLasso`	0.353	1.00/0.35	0.351	1.00/0.35	0.353	1.00/0.35
	`Lasso+Cov`	0.259	0.89/0.73	0.260	0.89/0.73	0.259	0.89/0.73
	`Lasso`	0.526	0.89/0.71	0.525	0.89/0.71	0.529	0.89/0.71

$p-$values < 0.01).

When $\mathbf{\Omega}$ is diagonal or dense, `GroupLasso+Cov` also attains the best estimation accuracy. Even though $\mathbf{\Omega}$ is not sparse in the latter setting, and our proposed estimator provides a sparse estimate of $\mathbf{\Omega}$, it still provides a considerable improvement over the `GroupLasso` by exploiting the correlated error term structure. Furthermore, the `GroupLasso+Cov` also significantly outperforms both lasso estimators.

Table 2.2 also contains the results on the True Positive Rate and True Negative Rate. The `GroupLasso+Cov` performs very similar to the `GroupLasso`. Accounting for the error correlation mainly affects the estimation accuracy, but only to a lesser extent the sparsity recognition performance. A similar observation is made by Rothman et al. [2010]. Furthermore, the grouplasso estimators attain, overall, a higher true positive rate than the lasso estimators.

Time series. First consider the settings with a sparse inverse error covariance structure. The MAEE for the VAR(2) model of dimension $q = 5$ is displayed

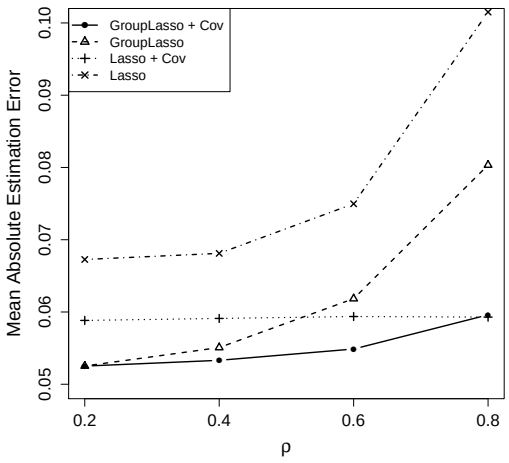


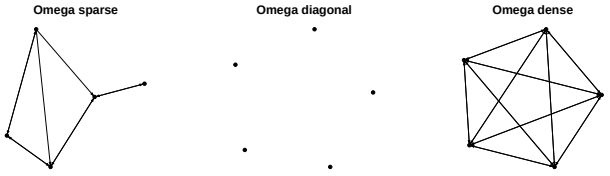
Figure 2.2: *VAR(2) model of dimension $q = 5$ and $T = 50$: Mean Absolute Estimation Error versus the correlation ρ, for the four considered estimators.*

in Figure 2.2 for different values of ρ. We find that (i) the improvement of `GroupLasso+Cov` over `GroupLasso` is remarkable when the error terms are highly correlated, (ii) `GroupLasso+Cov` and `GroupLasso` perform similarly when the error terms are hardly correlated, (iii) the grouplasso estimators perform, overall, better than the corresponding lasso estimators.

The MAEE for all simulation designs are reported in Table 2.3. For correlated errors (cfr. 'Omega sparse' and 'Omega dense'), the `GroupLasso+Cov` performs best and attains, in general, a considerably lower MAEE than the `GroupLasso`. For uncorrelated errors ('Omega diagonal'), the differences in estimation accuracy between `GroupLasso+Cov` and `GroupLasso` are less outspoken. Importantly, there is no loss in using the former compared to the latter. By sparsely estimating $\mathbf{\Omega}$, the absence of error correlation is accounted for.

Differences in the sparsity recognition between the estimators are less outspoken. While the estimators perform more similarly in terms of sparsity recognition, the considerable improvement in estimation accuracy attained by the `GroupLasso+Cov` gives it a clear advantage over the other estimators.

Table 2.3: *VAR(2) of dimension $q \in \{5, 20, 50\}$ and $T = 50$: Mean Absolute Estimation Error, True Positive and True Negative Rate.*



		$\rho = 0.6$					
	Estimator	MAEE	TPR/TNR	MAEE	TPR/TNR	MAEE	TPR/TNR
$q = 5$	`GroupLasso+Cov`	0.055	0.86/0.77	0.053	0.87/0.89	0.058	0.85/0.66
	`GroupLasso`	0.062	0.87/0.75	0.051	0.87/0.89	0.072	0.86/0.64
	`Lasso+Cov`	0.059	0.79/0.62	0.059	0.80/0.69	0.059	0.78/0.55
	`Lasso`	0.075	0.54/0.92	0.068	0.49/0.97	0.090	0.55/0.86
$q = 20$	`GroupLasso+Cov`	0.015	0.86/0.64	0.015	0.83/0.76	0.017	0.89/0.49
	`GroupLasso`	0.024	0.87/0.54	0.018	0.84/0.71	0.044	0.90/0.36
	`Lasso+Cov`	0.015	0.78/0.51	0.015	0.76/0.61	0.016	0.80/0.42
	`Lasso`	0.028	0.52/0.84	0.019	0.47/0.91	0.069	0.58/0.71
$q = 50$	`GroupLasso+Cov`	0.006	0.68/0.92	0.006	0.67/0.95	0.008	0.73/0.80
	`GroupLasso`	0.007	0.68/0.92	0.006	0.67/0.95	0.019	0.73/0.80
	`Lasso+Cov`	0.006	0.61/0.36	0.006	0.62/0.84	0.007	0.62/0.76
	`Lasso`	0.007	0.83/0.98	0.006	0.34/0.98	0.027	0.43/0.92

2.4 Application

We consider a data set of 30 mammary gland gene expression variables of mice [Abegaz and Wit, 2013]. Data are available for 18 time points, so we estimate a VAR(2) model of dimension $q = 30$, with $T = 18$. Since three samples are available, we estimate the VAR model three times.

We make an out-of-sample forecast comparison between `GroupLasso+Cov`, `GroupLasso`, `Lasso+Cov`, and `Lasso`. We use an expanding window approach. For $t = 13, \ldots, T-1$, we estimate the VAR(2) model using time points one until t and compute the one-step-ahead forecast. We compare the performance of the different estimators using the Mean Absolute Forecast Error

$$\text{MAFE} = \frac{1}{5} \sum_{t=13}^{T-1} \frac{1}{q} \sum_{i=1}^{q} |y_{t+1}^{(i)} - \widehat{y}_{t+1}^{(i)}|, \tag{2.9}$$

where $\widehat{y}_{t+1}^{(i)}$ is the estimate of the i^{th} response at time $t+1$. We repeat this exercise three times, once for each replicate sample. Results are given in Table 2.4.

Table 2.4: *Mean Absolute Forecast Error for the four considered estimators (rows) and three samples (column). The average MAFE, averaged over the three samples, is provided in the last column.*

Estimator	Sample 1	Sample 2	Sample 3	Average
`GroupLasso+Cov`	0.81	0.80	0.80	0.80
`GroupLasso`	1.23	1.38	1.72	1.44
`Lasso+Cov`	0.83	0.81	0.97	0.87
`Lasso`	1.51	1.85	2.37	1.91

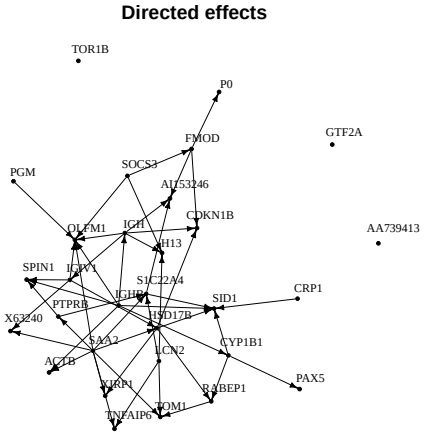


Figure 2.3: *Directed effects: a directed edge is drawn from one gene to another if the* `GroupLasso+Cov` *estimator indicates, by giving a non-zero regression estimate, that the former influences the latter.*

The `GroupLasso+Cov` attains the best forecast performance. It is closely followed by the `Lasso+Cov`. An important gain in prediction accuracy is obtained by accounting for the correlation structure of the error terms: the MAFE of the `GroupLasso+Cov` is, on average, 45% lower than the MAFE of the `GroupLasso`. Furthermore, we see from Table 2.4 that the grouplasso estimators perform better than the corresponding lasso estimators.

We study the interaction between the genes that trigger transitions to the mammary gland's main development stages. Figure 2.3 represents the 'directed, lagged effects' [Abegaz and Wit, 2013] inferred from $\widehat{\mathbf{B}}$. We discuss the results obtained from the first sample. Results for the other two samples are similar

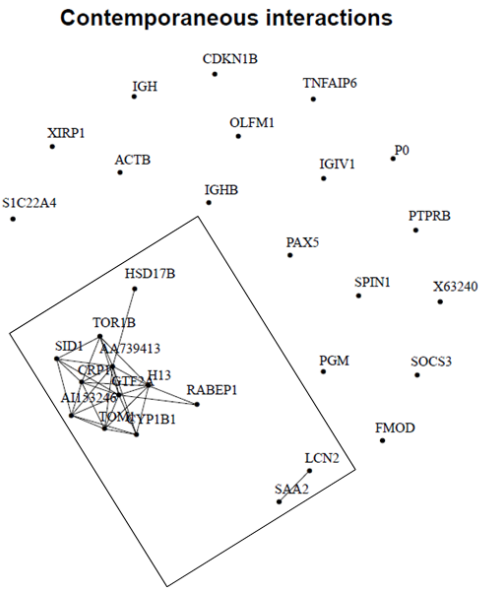


Figure 2.4: *Contemporaneous interactions: an undirected edge is drawn between two genes if the* `GroupLasso+Cov` *estimator indicates, by giving a non-zero estimate in* $\mathbf{\Omega}$*, that the innovations are partially correlated. Contemporaneous interactions are observed for only a subset of 13 genes, as indicated by the rectangle.*

and available from the authors upon request. The nodes in the network are the different genes. A directed edge from gene A to gene B is drawn if the `GroupLasso+Cov` indicates, by giving a non-zero estimate, that gene A has a lagged effect on gene B. The solution is very sparse: 850 out of the possible $900 = 30^2$ effects are estimated as zero. Some genes such as `GTF2A` and `TOR1B`, neither influence any other genes, nor are influenced by other genes. Other genes, such as `HSD17B` and `SAA2` are important hubs in the gene regulatory network. Previous research (Abegaz and Wit, 2013 and references therein) found these genes to play a central role in the mammary gland's development stages.

Figure 2.4 represents the 'contemporaneous interactions' [Abegaz and Wit, 2013] inferred from $\widehat{\mathbf{\Omega}}$. Again, the genes are the different nodes in the network. The elements of $\widehat{\mathbf{\Omega}}$ have a natural interpretation as partial correlations between the innovations (or error components) of the q equations in the VAR model. An edge is drawn between gene A and gene B if the corresponding element in the inverse error covariance matrix is estimated as non-zero. This means that the innovations

of genes A and B are contemporaneously partially correlated: conditional on all other innovations, a shock in the innovation of gene A will lead to an instantaneous shock in the innovation of gene B, and vice versa. As can be seen from Figure 2.4, contemporaneous interactions are observed only between a subset of 13 gene innovations, indicated by the rectangle. An important advantage of the sparse estimator is that the main interactions in the large gene regulatory network are highlighted. Out of the possible 435 interactions, only 32 are estimated as non-zero. As such, the researcher can concentrate on these results to further deepen our knowledge into the interactions at play in the development stages of the mammary gland.

Chapter 3

The predictive power of the business and bank sentiment of firms: A high-dimensional Granger Causality approach

Abstract

We study the predictive power of industry-specific economic sentiment indicators for future macro-economic developments. In addition to the sentiment of firms towards their own business situation, we study their sentiment with respect to the banking sector – their main credit providers. The use of industry-specific sentiment indicators results in a high-dimensional forecasting problem. To identify the most predictive industries, we present a bootstrap Granger Causality test based on the Adaptive Lasso. This test is more powerful than the standard Wald test in such high-dimensional settings. Forecast accuracy is improved by using only the most predictive industries rather than all industries.

3.1 Introduction

Sentiment indicators are often considered to be among the most important leading indicators of the real economy [Dreger and Kholodilin, 2013] and are therefore closely followed by business cycle analysts, central banks and business owners

(Vuchelen, 2004, Claveria et al., 2007, Martinsen et al., 2014). However, studies on the predictive power of sentiment indicators find mixed results. While many studies find that sentiment indicators have predictive power for future economic developments (Kumar et al., 1995, Hansson et al., 2005, Lemmens et al., 2005, Abberger, 2007, Klein and Oezmucur, 2010, Christiansen et al., 2014), others conclude that sentiment indicators provide only limited information for predicting economic variables (Cotsomitis and Kwan, 2006, Claveria et al., 2007, Dreger and Kholodilin, 2013 and Bruno, 2014).

An important communality between these studies is the use of aggregate sentiment indicators. This paper, instead, examines the predictive power of disaggregate sentiment indicators. Especially in the context of business sentiment – as is the topic of this paper – some segments have more predictive power than others. Here, we segment firms according to their industry. Our methodology takes into account that different industry segments might contain predictive power for different macro-economic indicators.

To study the predictive power, we use a Granger Causality approach. A set of time series Granger Causes another time series if the former has incremental predictive power for the latter. Granger Causality tests in *low-dimensional* time series settings have a long history. They are used, among others, in macroeconomics to study the predictive power of monetary aggregates for output and price variables [Sahoo and Acharya, 2010], in operational research to study the predictive power of academic literature for practitioner literature [Ghosh et al., 2010], or in finance to study the predictive power of volume for stock prices [Blasco et al., 2005]. Because predictive analysis based on disaggregate sentiment indicators requires handling a large number of such indicators, we introduce a Granger Causality test for *high-dimensional* time series data.

Recently, testing procedures for high-dimensional cross-section data have gained attention, for instance Wasserman and Roeder [2009], Meinshausen et al. [2009] and Chatterjee and Lahiri [2011]. We extend the residual bootstrap procedure of Chatterjee and Lahiri [2011] to high-dimensional *time series* data. The bootstrap test statistic, based on the Adaptive Lasso [Zou, 2006], identifies those industry segments whose predictive power is statistically significant. Our simulation study shows that this test statistic is more powerful than the standard Wald test statistic in a high-dimensional setting. Furthermore, important gains in forecast accuracy are obtained by not using all industry segments but by first selecting the most predictive ones using the bootstrap test.

We use a unique data set that not only measures the sentiment of firms towards

their own situation ('*business* sentiment') – as is classical for sentiment indicators – but also measures the sentiment of firms towards the banking industry ('*bank* sentiment'). For the economy to be able to grow, it is essential that firms have access to credit, typically provided by banks. Especially in the aftermath of the recent economic downturn and banking crises, distressed banks can constrain the economy (Kroszner et al., 2007, Dell'Ariccia et al., 2008, Fernandez et al., 2013).

The remainder of this article is structured as follows. In Section 3.2, we discuss the contribution of our paper to the Business Sentiment literature. Section 3.3 describes the data on business and bank sentiment, as well as the macro-economic indicators. Section 3.4 introduces Granger Causality Testing in high-dimensional time series models. In Section 3.5, a simulation study shows the good performance of our methodology in terms of size and power of the test statistic and forecast accuracy. In Section 3.6, we apply the proposed methodology to identify the most predictive industry segments for several future macro-economic indicators. In Section 3.7, we show that forecast accuracy can be improved by using only the most predictive industry segments instead of all industry segments. The robustness of our findings is investigated in Section 3.8. Finally, Section 3.9 concludes.

3.2 Contribution

Our objective is to study the predictive power of Business Sentiment Surveys for future macro-economic growth. Business Sentiment Surveys are carried out on a monthly basis by various public and private institutions. These surveys are the most popular channel to get insight into the beliefs of economic agents at the supply side of the economy. If business owners feel confident about their current and future economic situation, they might invest more and increase their activity. Hence, Business Sentiment Surveys are often seen as early indicators for future economic developments.

The Joint Harmonized EU Programme of Business and Consumer Sentiment Surveys systematically collects sentiment data using surveys. The Business Sentiment Survey includes questions on several aspects of the firm's economic situation, such as their expected production, selling prices and exports. In contrast to Consumer Sentiment Surveys that include questions on the consumer's assessment of the overall economy, Business Sentiment Surveys typically only consist of an evaluation of each firm's own economic situation, i.e. the well-known 'business sentiment' (e.g. Hansson et al., 2005, Lemmens et al., 2005, Abberger, 2007,

Claveria et al., 2007, Klein and Oezmucur, 2010, Gelper and Croux, 2010, Christiansen et al., 2014.).

In addition to business sentiment, we also study 'bank sentiment', i.e. the sentiment of firms towards the banking industry. Studying bank sentiment is relevant since access to financial resources is crucial for firms being able to grow. Typically, these financial resources are provided by banks. This is especially true for small- to medium-sized firms (e.g. Beck and Demirguc-Kunt, 2006, Angilella and Mazzu, 2015). Germany, the country we study in this paper, is dominated by this type of companies: in our sample, around 93% of the respondents are small- to medium-sized firms. To the best of our knowledge, we are the first to study the importance of sentiment towards the banking industry.

Studying the predictive power of these business and bank sentiment indicators is challenging given the large amount of sentiment indicators that is available. In our sentiment application, 150 sentiment indicators are measured over 40 months. We combine all 150 sentiment indicators in one large model. To handle this high-dimensionality, we use penalized maximum likelihood estimation. Our approach also involves a selection procedure: out of the 150 sentiment indicators, we select the most predictive ones using a Granger Causality test. These selected sentiment indicators are then used to forecast macro-economic growth.

To handle the high-dimensionality of sentiment data, previous studies either (i) summarize the information from all individual sentiment indicators into a aggregated sentiment indicator and study the latter's predictive power (Hansson et al., 2005, Abberger, 2007, Claveria et al., 2007, Klein and Oezmucur, 2010, Christiansen et al., 2014, Gelper and Croux, 2010), or (ii) estimate separate models for the individual sentiment indicators and combine the forecast from these models [Martinsen et al., 2014]. However, these approaches involve several issues. By aggregating, one risks losing valuable information. Though aggregate indicators are often followed by business analysts and used in economic research, the individual sentiment indicators might contain even more relevant and interesting information (Roos, 2008). Indeed, Martinsen et al. [2014] find that forecast models with individual sentiment indicators considerably improve models with aggregated sentiment indicators. An advantage of our approach compared to the forecast combination approach of Martinsen et al. [2014] is that we investigate whether forecast performance can be improved by using only the most predictive indicators instead of using all. Our empirical results, to be discussed in Section 3.7, show that further improvements in forecast performance are indeed obtained by using only the most predictive indicators.

Table 3.1: *Industry Segments. Businesses are divided into 10 industry segments.*

Industry	Description	Sector
Industry 1	Agriculture, forestry, fishing, mining and quarrying and other industry	Primary
Industry 2	Manufacturing	Secondary
Industry 3	Construction	Secondary
Industry 4	Wholesale and retail trade, transportation and storage accommodation and food and service activities	Tertiary
Industry 5	Information and communication	Quaternary
Industry 6	Financial and insurance activities	Quaternary
Industry 7	Real estate activities	Quaternary
Industry 8	Professional, scientific, technical administration and support service activities	Quaternary
Industry 9	Public administration, defense, education	Quaternary
Industry 10	Other services	Quaternary

3.3 Data

We use a unique data set provided to us by EUWIFO, the European Economic Research Institute. EUWIFO is an owner-managed business that conducts business climate interviews. By conducting interviews with firms spread over Germany, EUWIFO gathers information on the confidence these firms have in their own economic situation and in the banking sector. Firms are divided into segments according to the industry in which they are active. To this end, we use NACE codes since this is the standard business classification framework in the European Union (e.g. Weinstein, 2013). We consider 10 industry segments, as listed in Table 3.1.

The interviews consist of two parts. In the first part, the Business Survey, firms are asked to assess their own situation. In the second part, the Bank Survey, firms are asked to assess the German bank sector.

Business Survey Each firm receives 9 questions to assess their own economic situation. They are asked to assess changes (this year compared to last year) in (1) turnover, (2) earnings, (3) number of employees, (4) investments, (5) incoming domestic orders, (6) incoming foreign orders, (7) utility and maintenance costs, (8) tax burden, and (9) cost through government red tape. For each question, answers are favorable, neutral or unfavorable. For all the firms within an industry segment, we calculate a balance of opinion for each question, defined as the percentage of favorable answers minus the percentage of unfavorable answers. We construct 9 such sentiment indicators for each of the 10 industries, which amounts to 90 business sentiment indicators.

Table 3.2: *Macro-economic indicators. All time series are seasonally adjusted (Eurostat).*

Indicator	Description
IP-A1	Production in industry: Mining and quarrying; manufacturing; electricity, gas, steam and air conditioning supply
IP-A2	Production in industry: Construction, Mining and quarrying; manufacturing; and electricity, gas, steam air conditioning supply
IP-M	Production in industry: Manufacturing
IP-E	Production in industry: Energy
IP-CaGo	Production in industry: Capital goods
IP-CoGo	Production in industry: Consumer goods
RT	Retail Trade, except of motor vehicles and motorcycles
WS	Wholesale Trade, except of motor vehicles and motorcycles

Bank Survey Each firm is asked to assess the German bank sector. In total, 243 German banks are included in the Bank Survey. Each firm first has to indicate which of these 243 German banks they know. For the banks they know, they are asked to assess their *consideration* towards that specific bank and the *reputation* of that specific bank. Answers are either favorable or unfavorable and a balance of opinion indicator is calculated for each question. We include three indicators: the average consideration indicator, averaged over all German banks, the consideration indicator towards the Sparkassen, and the consideration indicator towards the Volksbanken. The latter two are the most well known banks in Germany. We also construct three reputation indicators per industry segment following an analogous approach. As we construct three bank consideration and three bank reputation indicators for each of the 10 industries, this amounts to 60 bank sentiment indicators.

Joining the 90 business sentiment indicators and the 60 bank sentiment indicators results in a total of 150 time series. We combine all 150 sentiment indicators in one high-dimensional data set. All time series are observed over $T = 40$ months (January 2012-April 2015). We study the predictive power of these sentiment indicators for 8 German macro-economic indicators (Table 3.2).

The 150 time series are grouped into blocks by industry segment (cfr. Table 3.1). For each of the 10 industry segments, we have one block of 9 indicators from the Business Survey and one block of 6 indicators from the Bank Survey. Our methodology is such that we select either all 9 business sentiment indicators for an industry, or none. Similarly, we will select either all 6 bank sentiment indicators for an industry or none. This way, we can investigate the difference in predictive

power between the business and bank sentiment indicators. To identify the most predictive blocks, we perform joint hypothesis tests. We test if the set of indicators in a particular block Granger Causes a particular macro-economic indicator. This predictive analysis involves a large number of disaggregate sentiment indicators. In the next section, we introduce a Granger Causality testing procedure that can handle such a high-dimensional situation.

3.4 High-dimensional Granger Causality Testing

Performing Granger Causality tests on a data set with many time series relative to the length of the series is challenging. In these high-dimensional settings, estimation by standard procedures becomes inaccurate. In our sentiment application, the number of time series (i.e. $k = 150$) even exceeds the length of the time series (i.e. 40), making it impossible to use standard estimation procedures. Penalized estimation brings an outcome.

3.4.1 Penalized Maximum Likelihood estimation

Let y_t be a one-dimensional stationary time series. We assume that y_t follows a ARX(p) model, i.e. an autoregressive model of order p with k predictor time series collected in the $(k \times 1)$ vector $\mathbf{x}_t$:

$$y_t = b_1 y_{t-1} + b_2 y_{t-2} + \ldots + b_p y_{t-p} + \mathbf{a}_1 \mathbf{x}_{t-1} + \mathbf{a}_2 \mathbf{x}_{t-2} + \ldots + \mathbf{a}_p \mathbf{x}_{t-p} + e_t \,, \quad (3.1)$$

where b_1 to b_p are the autoregressive parameters, the parameters $\mathbf{a}_1$ to $\mathbf{a}_p$ are $(1 \times k)$ vectors and the error term e_t is assumed to follow a $N(0, \sigma)$ distribution. We assume, without loss of generality, that all time series are mean centered such that no intercept is included.

If the number of components in $\mathbf{x}_t$ is large, the number of unknown parameters in equation (3.1) explodes. To ensure accurate estimation, we use Penalized Maximum Likelihood estimation (e.g. Zou, 2006 in a regression context, or Gelper et al., 2016 in a time series context). Write model (3.1) in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \,, \quad (3.2)$$

where $\mathbf{y}$ is the column vector $(y_1, \ldots, y_T)$, and the matrix $\mathbf{X} = (\underline{\mathbf{Y}}_1, \ldots, \underline{\mathbf{Y}}_p, \underline{\mathbf{X}}_1, \ldots, \underline{\mathbf{X}}_p)$. Here $\underline{\mathbf{Y}}_j$ is $(T \times 1)$, containing the values of the time series at lag

j in its column; and $\underline{\mathbf{X}}_j$ is an $(T \times k)$ matrix, containing the values of the k predictor time series at lag j in its columns, for $1 \leq j \leq p$. The vector $\boldsymbol{\beta}$ contains the parameters values $b_1, \ldots, b_p, \mathbf{a}_1, \ldots, \mathbf{a}_p$, and has length $p(1+k)$. In case $p(1+k) > T$, the Maximum Likelihood estimator does not exist. The Penalized Maximum Likelihood estimator is, however, still computable.

The penalized estimator of the regression parameter $\boldsymbol{\beta}$ is obtained by minimizing the negative log likelihood with a penalization on the elements of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p(1+k)} \hat{w}_i |\beta_i| \,, \tag{3.3}$$

where $\hat{w}_i$ are weights and $\lambda > 0$ is a sparsity parameter. This estimator is the Adaptive Lasso [Zou, 2006]. It generalizes the popular Lasso (e.g. Hastie et al., 2009, Chapter 3) which shows good performance in operational research (e.g. Ballings and Van den Poel, 2015, Huang et al., 2014). Use of the Adaptive Lasso ensures that the bootstrap (Section 3.4.3) is consistent [Chatterjee and Lahiri, 2011]. We take the weights of the Adaptive Lasso as $\hat{w}_i = 1/|\hat{\beta}_i^{\text{ridge}}|$, where the Ridge estimator (Hastie et al., 2009, Chapter 3) is

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_{\text{ridge}} \sum_{i=1}^{p(1+k)} \beta_i^2.$$

The sparsity parameter λ and the order of the ARX, p, are selected using the Bayesian Information Criterion (BIC) (e.g. Abegaz and Wit, 2013 and references therein):

$$\text{BIC}_\lambda = T \cdot \log\left(\frac{1}{T}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda)^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda)\right) + df_\lambda \cdot \log(T),$$

where df_λ equals the number of non-zero estimated regression coefficients. We solve (3.3) over a range of values for λ and select the one with lowest value of the BIC. To select the order of the ARX model, we estimate the ARX model for different values of p, each time using the optimal value of λ for that value of p. We then select the order p of the ARX model again by minimizing the BIC.

3.4.2 Granger Causality in the ARX framework

We partition the vector $\mathbf{x}_t$ in different blocks, and denote the j^{th} block of $\mathbf{x}_t$ by $\mathbf{x}_{t,j}$, consisting of k_j time series. In the ARX model (3.1), denote the j^{th} block of coefficients at lag i corresponding to $\mathbf{x}_{t,j}$ by $\mathbf{a}_{i,j}$. The multivariate time series

$\mathbf{x}_{t,j}$ is said to Granger Cause y_t if the former has incremental predictive power for the latter. We say that $\mathbf{x}_{t,j}$ does not Granger Cause y_t if the coefficients on all lags of $\mathbf{x}_{t,j}$ are equal to zero, i.e. $\mathbf{a}_{1,j} = \ldots = \mathbf{a}_{p,j} = \mathbf{0}$.

The Adaptive Lasso estimator in (3.3) is sparse, meaning that some of its elements are exactly zero. The larger the value of λ, the sparser the estimator. The 'Granger Lasso Selection' method (e.g. Fujita et al., 2007, Bahadori and Liu, 2013) says that a time series $\mathbf{x}_{t,j}$ Granger Causes y_t if at least one of the corresponding parameters $\mathbf{a}_{1,j}, \ldots, \mathbf{a}_{p,j}$ is estimated as non-zero. Our approach is different, we infer Granger Causality relations from a bootstrap testing procedure.

3.4.3 Granger Lasso test

The null hypothesis that a block of time series $\mathbf{x}_{t,j}$ is not Granger Causing y_t can be stated as

$$H_0 : \ \mathbf{R}_j \boldsymbol{\beta} = \mathbf{0}, \tag{3.4}$$

where $\mathbf{R}_j$ is a suitable $pk_j \times p(1+k)$ matrix. The elements of $\mathbf{R}_j$ are either zero or one. We assign the value one to the elements of $\mathbf{R}_j$ corresponding to the autoregressive parameters $\mathbf{a}_{1,j}, \ldots, \mathbf{a}_{p,j}$. The corresponding Wald test statistic is given by

$$Q = (\mathbf{R}_j \widehat{\boldsymbol{\beta}})^T (\mathbf{R}_j \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) \mathbf{R}_j^T)^{-1} (\mathbf{R}_j \widehat{\boldsymbol{\beta}}). \tag{3.5}$$

To bootstrap this test statistic, we use the following residual bootstrap procedure (Kreiss and Lahiri, 2012):

(i) Estimate the model under the null hypothesis, i.e. model (3.1) with the block $\mathbf{x}_{t,j}$ removed at the right-hand-side. Compute the centered residuals $\widehat{\varepsilon}_t$, for $t = 1, \ldots, T$.

(ii) Let $B = 500$ be the number of bootstraps. For $b = 1, \ldots, B$:

  (a) Construct the bootstrap time series y_t^* from model (3.1) with the parameter estimates from step 1 and with bootstrap errors $\varepsilon_t^* = \widehat{\varepsilon}_{\mathcal{U}_t}$ with $\mathcal{U}_t, t = 1, \ldots, T$ an i.i.d. sequence of discrete random variables uniformly distributed on $\{1, \ldots, T\}$.[1] The predictor time series are kept fixed.

  (b) Apply the Penalized Maximum Likelihood estimator of equation (3.3) to the bootstrap sample. Denote the bootstrap estimate by $\widehat{\boldsymbol{\beta}}_b^*$.

[1] We check if the bootstrap errors are white noise using the Ljung-Box test. If so, we continue to the next step, otherwise we re-draw from the centered residuals.

Table 3.3: *Simulation designs.*

Design	under H_0	under H_A
$T = 100, k = 25$	$\mathbf{a}_1 = \begin{bmatrix} \mathbf{0.2}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times(k-15)} \end{bmatrix}$	$\mathbf{a}_1 = \begin{bmatrix} \mathbf{0.2}_{1\times 5} & \mathbf{0.2}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times(k-15)} \end{bmatrix}$
$T = 100, k = 50$	$\mathbf{a}_1 = \begin{bmatrix} \mathbf{0.2}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times(k-15)} \end{bmatrix}$	$\mathbf{a}_1 = \begin{bmatrix} \mathbf{0.2}_{1\times 5} & \mathbf{0.2}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times(k-15)} \end{bmatrix}$
$T = 100, k = 75$	$\mathbf{a}_1 = \begin{bmatrix} \mathbf{0.2}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times(k-15)} \end{bmatrix}$	$\mathbf{a}_1 = \begin{bmatrix} \mathbf{0.2}_{1\times 5} & \mathbf{0.2}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times(k-15)} \end{bmatrix}$
$T = 40, k = 150$	$\mathbf{a}_1 = \begin{bmatrix} \mathbf{0.4}_{1\times 9} & \mathbf{0}_{1\times 9} & \dots & \mathbf{0}_{1\times 9} & \mathbf{0}_{1\times 6} & \dots & \mathbf{0}_{1\times 6} \end{bmatrix}$	$\mathbf{a}_1 = \begin{bmatrix} \mathbf{0.4}_{1\times 9} & \mathbf{0.4}_{1\times 9} & \mathbf{0}_{1\times 9} & \dots & \mathbf{0}_{1\times 9} & \mathbf{0}_{1\times 6} & \dots & \mathbf{0}_{1\times 6} \end{bmatrix}$

(c) Compute the bootstrap statistic $Q_b^* = (\mathbf{R}_j\widehat{\boldsymbol{\beta}}_b^*)^T(\mathbf{R}_j\text{Cov}(\widehat{\boldsymbol{\beta}})\mathbf{R}_j^T)^{-1}(\mathbf{R}_j\widehat{\boldsymbol{\beta}}_b^*)$.

(iii) Compute

$$\text{mid } p\text{-value } = \frac{1}{B}\sum_{b=1}^{B}\Big(I(Q_b^* > Q) + \frac{1}{2}I(Q_b^* = Q)\Big),$$

with Q_b^* (for $b = 1, \dots, B$) B independent bootstrap statistics. $I(\cdot)$ is an indicator function that takes on the value one if its argument is true and equals zero otherwise. We use the mid p-value [Lancaster, 1949] since it may occur that the value of the test statistic and the bootstrap test statistic are both equal to zero.

3.5 Simulation study

By means of a simulation experiment, we (i) evaluate the size and power of the Granger Lasso test and (ii) conduct a forecast exercise. We generate y_t according to the following ARX(1) model

$$y_t = 0.5y_{t-1} + \mathbf{a}_1\mathbf{x}_{t-1} + e_t, \tag{3.6}$$

where $e_t \sim N(0, 0.1)$. The predictors are generated as autoregressive processes $\mathbf{x}_t = \mathbf{C}\mathbf{x}_{t-1} + \mathbf{u}_t$, with $\mathbf{u}_t \sim N_k(\mathbf{0}, 0.1\mathbf{I})$, $\mathbf{C} = 0.5\mathbf{I}$ and $\mathbf{I}$ the k-dimensional identity matrix. The model parameters are chosen according to the four designs detailed in Table 3.3. The first three designs are the same except for the number of time series k. In design two and three, we add more non-informative time series to the model, i.e. time series with a coefficient equal to zero. The standard Maximum Likelihood estimator is computable in these three designs. The last design corresponds to the design of our sentiment application, with $k = 150$ predictor time series and $T = 40$. Here, only the Penalized Maximum Likelihood estimator is computable.

For each design, we consider a data generating process under the null hypothesis H_0 and under the alternative hypothesis H_A. We divide the time series $\mathbf{x}_t$ and the corresponding coefficient vector $\mathbf{a}_1$ into several blocks, as can be seen from Table 3.3. The first block of time series Granger Cause the response both under H_0 and under H_A. The second block of time series Granger Cause the response only under H_A. The remaining blocks of time series never Granger Cause the response. In the first three designs, block one to three each contain five time series, the fourth block contains the remaining ones. In the last design, there are 20 blocks, similar to our sentiment application.

3.5.1 Size and power of the test statistic

We test the null hypothesis that the second block of time series does not Granger Cause the response. We compare the performance of Granger Lasso test to the standard Wald test computed from the standard Maximum Likelihood (ML) estimator.

To study the *size* of the test statistic, we simulate $N = 1000$ time series under the null hypothesis and compute the simulated size, i.e. the proportion of simulation runs were the null hypothesis is rejected:

$$\text{Simulated size} = \frac{1}{N}\sum_{j=1}^{N} I(p_j^{H_0} < \alpha), \tag{3.7}$$

where $p_j^{H_0}$ is the mid p-value obtained in simulation run $j = 1, \ldots, N$, and α is the pre-specified significance level. We consider $\alpha = 0.01$ and $\alpha = 0.05$.

Results. Table 3.4 shows the simulated sizes for the standard Wald test and the Granger Lasso test. The simulated sizes of the Granger Lasso test and the standard Wald test are both close to the nominal size α in the design with $T = 100, k = 25$. When the number of time series increases relative to the length of the time series (i.e. second and third design), the Granger Lasso test remains accurately sized whereas the standard Wald test statistic gets distorted: its simulated size deviates strongly from the nominal size. In the last design, only the Granger Lasso test is available. For both $\alpha = 0.01$ and $\alpha = 0.05$, the Granger Lasso test is reasonably accurately sized.

To study the *power* of the test statistic, we use size-power curves (see Davidson and McKinnon, 1998). Size-power curves are constructed using two empirical distribution functions. We carry out the following steps:

Table 3.4: *Simulated sizes for the Wald test and Granger Lasso test.*

Simulation design	Wald test		Granger Lasso test	
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
$T = 100, k = 25$	0.017	0.064	0.013	0.058
$T = 100, k = 50$	0.025	0.079	0.010	0.052
$T = 100, k = 75$	0.035	0.082	0.015	0.051
$T = 40, k = 150$	NA	NA	0.007	0.051

(i) Simulate $N = 1000$ time series under the null hypothesis. Compute for each simulation run $j = 1, \ldots, N$ the mid p-value $p_j^{H_0}$. Calculate the empirical distribution function of the p-values:

$$\widehat{F}^{H_0}(x_i) = \frac{1}{N} \sum_{j=1}^{N} I(p_j^{H_0} \leq x_i),$$

for a grid of values $x_i, i = 1, \ldots, m$ between zero and one.

(ii) Simulate $N = 1000$ time series under the alternative hypothesis. Compute for each simulation run $j = 1, \ldots, N$ the mid p-value $p_j^{H_A}$. Calculate

$$\widehat{F}^{H_A}(x_i) = \frac{1}{N} \sum_{j=1}^{N} I(p_j^{H_A} \leq x_i).$$

(iii) Plot $\widehat{F}^{H_0}(x_i)$ against $\widehat{F}^{H_A}(x_i)$, for $x_i, i = 1, \ldots, m$.

Results. Size-power curves of the Granger Lasso test and standard Wald test are shown in Figure 3.1 (first three designs). The larger the difference between the size-power curve and the 45°line, the more power the test has. For $k = 25$ (i.e. left panel) both curves are rapidly increasing and very similar. When the number of time series increases (i.e. middle and right panel), the size-power curve of the Granger Lasso test is hardly affected, and achieves a much larger power than the standard Wald test.

3.5.2 Forecasting

For forecasting the time series y_t, we use a two-step procedure. First, we select predictor time series. Second, we estimate the model with only the selected predictor time series. We consider four selection and six estimation techniques,

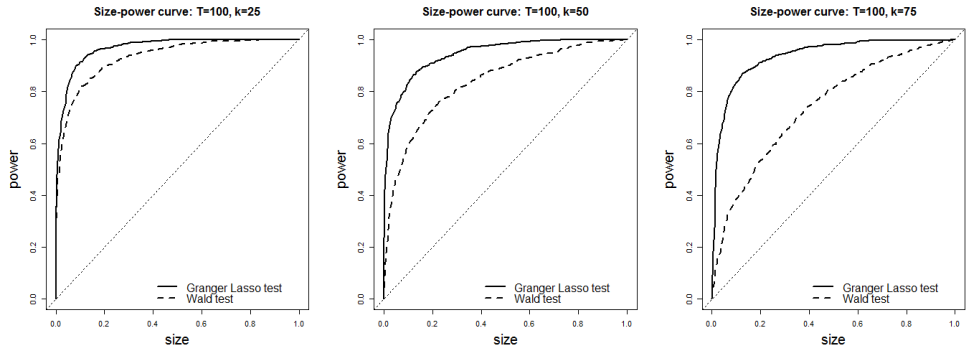


Figure 3.1: *Size-power curve of the Granger Lasso test (solid line) and the standard Wald test (dashed line), for increasing number of time series $k = 25$ (left), $k = 50$ (middle) and $k = 75$ (right) with time series length $T = 100$. The 45° line (dotted line) is indicated as well.*

yielding 24 selection-estimation combinations. We investigate the performance of each combination in forecasting the response.

As selection techniques we consider: (1) use all time series, (2) use the standard Wald test to discard blocks of time series that are not Granger Causing the response, (3) use Granger Lasso Selection (cfr. Section 3.4.1) to discard blocks of time series that are not Granger Causing the response, (4) use the Granger Lasso test to discard blocks of time series that are not Granger Causing the response. Selection technique (4) is our proposed selection technique. The tests are carried out at a 1% significance level.

After selecting the predictor time series, we forecast the response using either (1) Maximum Likelihood, (2) the Adaptive Lasso estimator, (3) Bayesian shrinkage with the Minnesota prior [Litterman, 1986], (4) the Factor Model of Stock and Watson [2002], (5) Bagging (Hastie et al., 2009, Chapter 8), (6) Random Forest (Hastie et al., 2009, Chapter 15). These are all leading methods for macro-economic forecasting [Inoue and Kilian, 2008]. Methods (2) and (3) perform shrinkage. While the Adaptive Lasso puts some of the estimated coefficients exactly to zero, the Bayesian estimator only shrinks the estimated coefficients towards zero. Factor Models reduce the dimension of the predictor time series by extracting a small number of common factors using principal component analysis.[2]

[2] The number of factors r is determined by calculating the maximum eigenvalue ratio criterion $\hat{r}_j = \hat{\lambda}_j / \hat{\lambda}_{j+1}$ for $j = 1, \ldots, \mathrm{k}-1$ from the eigenvalues $\hat{\lambda}_j, \ldots, \hat{\lambda}_\mathrm{k}$ and selecting $r = \mathrm{argmax}_j \hat{r}_j$.

Bagging draws bootstrap samples from the original model, makes a prediction - for which we use the Adaptive Lasso - based on each bootstrap sample and averages these predictions over the bootstrap samples. We use a stationary bootstrap [Politis and Romano, 1994] to account for the time series structure of the data in the construction of the bootstrap samples.[3] Random Forest differs from Bagging in that it selects random subsets of predictors. We take this subset size to be equal to the square root of the number of predictors.

To evaluate forecast accuracy, we conduct a rolling window forecast exercise. We use a window of size $S = \lfloor 0.90 \cdot T \rfloor$. At each point $t = S, \ldots, T-1$, the models are re-estimated and one-step-ahead forecasts are calculated. We evaluate the forecast accuracy of each selection-estimation technique combination by calculating the Mean Absolute Forecast Error[4]

$$\text{MAFE} = \frac{1}{T-S} \sum_{t=S}^{T-1} \left| \hat{y}_{t+1} - y_{t+1} \right|, \tag{3.8}$$

where $\hat{y}_{t+1}$ is the predicted response for time $t+1$. The MAFE is computed for each simulated time series, and their average over $N = 100$ simulation runs is reported in Table 3.5.

Results. Table 3.5 shows that selecting predictor time series is better than taking all series, for all estimation techniques (except the Bayesian shrinkage estimator). Among the selection techniques, improvements are, overall, larger with the proposed Granger Lasso test compared to the Granger Lasso Selection approach. Granger Lasso Selection discards less blocks of time series compared to the Granger Lasso test, yielding less parsimonious models and reduced forecast performance. When the number of time series increases relative to the length of the time series, the Granger Lasso test also performs substantially better than the standard Wald test. Paired t-tests confirm that (in the large majority of cases), the improvements of the Granger Lasso test compared to the other selection techniques are significant. More precisely, the Granger lasso test performs significantly best - among the four selection techniques - in 15 out of 18 cases (design $T = 100, k = 50$), 16 out of 18 cases (design $T = 100, k = 75$), and 8 out of 10 cases (design $T = 40, k = 150$). The good performance of the Granger Lasso test is most pronounced in the high-dimensional designs.

[3] The block resampling uses a random bootstrap block size generated from a genometric distribution with mean six. We take $B = 100$ bootstrap samples.

[4] Similar conclusions can be drawn by looking at the Mean Squared Forecast Error.

Table 3.5: *Average MAFE for the four selection techniques (rows) and six estimation techniques (columns).*

Simulation design	Selection technique	Estimation technique					
		ML	Adaptive Lasso	Bayesian	Factor Model	Bagging	Random Forest
$T = 100, k = 25$	All	0.093	0.089	0.116	0.129	0.230	0.224
	Wald test	0.082	0.082	0.121	0.086	0.225	0.215
	Granger Lasso Selection	0.089	0.085	0.118	0.121	0.176	0.200
	Granger Lasso test	0.082	0.082	0.120	0.086	0.175	0.200
$T = 100, k = 50$	All	0.126	0.092	0.122	0.138	0.268	0.243
	Wald test	0.087	0.084	0.124	0.089	0.328	0.235
	Granger Lasso Selection	0.119	0.092	0.122	0.137	0.186	0.209
	Granger Lasso test	0.084	0.083	0.124	0.086	0.192	0.209
$T = 100, k = 75$	All	0.208	0.089	0.123	0.141	0.156	0.161
	Wald test	0.117	0.088	0.121	0.107	0.197	0.156
	Granger Lasso Selection	0.170	0.091	0.123	0.140	0.104	0.124
	Granger Lasso test	0.083	0.080	0.119	0.085	0.141	0.136
$T = 40, k = 150$	All	NA	0.189	0.315	0.322	0.648	0.535
	Granger Lasso Selection	NA	0.181	0.305	0.300	0.627	0.461
	Granger Lasso test	NA	0.165	0.379	0.199	0.472	0.385

For all simulation designs, the best forecast always includes the Granger Lasso test. Among the estimation techniques, the Adaptive Lasso performs best. After the first selection of predictive blocks of time series, the Adaptive Lasso can further reduce the number of predictor time series in the second step. This is most suited for settings with few relevant predictor time series and many irrelevant, noise predictor time series. Similar conclusions are obtained by Bühlmann and Hothorn [2010] who discuss a 'Twin Boosting' procedure for improved feature selection and prediction.

3.6 The role of business and bank sentiment for macro-economic forecasting

We identify the most predictive industry segments for future macro-economic developments using the Granger Lasso test from Section 3.4.

3.6.1 Model

We estimate 8 ARX models, one for each macro-economic indicator to predict. Following standard practice, we study the predictive power of sentiment *change* indicators for macro-economic *growth*. Hence, the time series y_t entering model (3.1) is one of the 8 macro-economic indicators of Table 3.2 taken in log-differences. The vector $\mathbf{x}_t$ contains the $k = 150$ business and bank sentiment indicators in first differences at time t. To ensure a uniform treatment of all time series, we check for stationarity using the pooled unit root test from Levin et al. [2002]. We find the 8 macro-economic growth time series to be jointly stationary (p-value< 0.01), as well as the 150 sentiment change time series (p-value< 0.01). The Augmented Dickey-Fuller tests also indicate each individual time series to be stationary.

We estimate each ARX model using the Penalized Maximum Likelihood estimator from Section 3.4. Then, we perform Granger Causality tests, one for each of the 20 blocks of sentiment indicators (cfr. Section 3.3). As such, we test if the change in opinion of a particular industry segment - as measured through the Business Survey - has incremental predictive power for the German macro-economic growth indicators. We repeat this exercise for each industry segment using the Bank Survey.

3.6.2 Identifying the most predictive industries

For each industry, Table 3.6 reports the p-value of the test that the change in opinion of that particular industry does not Granger Cause a particular macro-economic growth indicator. Significant results at the 1% level are in bold. We discuss the results by building on the sectoral classification framework which distinguishes the primary, secondary, tertiary and quaternary sector.

Business Survey. The primary sector, unlike the other sectors, has almost no incremental predictive power. The primary sector's contribution to Germany's GDP is also the smallest. The secondary industry has most incremental predictive power for the macro-economic growth indicators to which these sectors contribute most (IP-A1, IP-A2, IP-M and IP-E). Firms active in the tertiary and especially the quaternary sector have incremental predictive power for several macro-economic growth indicators. This sector consists of the knowledge-based part of the economy, and accounts for roughly 65% of Germany's GDP. Firms active in these sectors are at the heart of the whole economy.

Bank Survey. The Bank Survey contains less incremental predictive power

Table 3.6: *P-values of the Granger Causality test with null hypothesis that the change in opinion of a particular industry segment (rows) does not Granger Cause a particular macro-economic growth indicator (columns). Significant results at the 1% level are in bold.*

			Macro-economic indicators							
	Industry segment	Sector	IP-A1	IP-A2	IP-M	IP-E	IP-CaG	IP-CoG	RT	WS
Business Survey	Agriculture, mining & other industry	Primary	0.02	0.03	0.02	0.04	**0.00**	0.01	0.08	0.25
	Manufacturing	Secondary	**0.00**	0.05	**0.01**	**0.00**	**0.00**	**0.00**	0.23	0.09
	Construction	Secondary	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.64	0.16	0.12
	Wholesale, retail trade, transportation, food & service	Tertiary	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.87	0.21	**0.00**
	Information & communication	Quaternary	0.91	0.14	0.88	0.96	0.98	0.90	0.66	**0.00**
	Finance	Quaternary	0.36	**0.00**	0.10	**0.00**	**0.00**	0.67	**0.00**	0.74
	Real estate	Quaternary	0.82	0.80	0.72	**0.00**	1.00	**0.00**	**0.00**	0.60
	Administration & support	Quaternary	0.05	0.12	**0.00**	**0.00**	**0.00**	**0.00**	0.73	**0.00**
	Public services	Quaternary	0.80	0.04	0.82	0.04	**0.00**	**0.00**	0.71	0.64
	Other services	Quaternary	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.66	0.60	**0.01**
Bank Survey	Agriculture, mining & other industry	Primary	1.00	1.00	1.00	1.00	1.00	0.99	0.97	0.99
	Manufacturing	Secondary	0.12	0.91	0.18	1.00	1.00	0.96	0.52	0.86
	Construction	Secondary	0.94	1.00	1.00	0.06	1.00	0.01	0.80	0.14
	Wholesale, retail trade, transportation, food & service	Tertiary	1.00	0.63	1.00	1.00	1.00	0.03	0.06	0.69
	Information & communication	Quaternary	0.12	**0.00**	0.01	1.00	0.01	0.95	**0.00**	0.70
	Finance	Quaternary	0.97	0.94	0.98	**0.00**	1.00	0.03	0.12	**0.00**
	Real estate	Quaternary	0.78	0.84	0.41	1.00	1.00	0.93	0.79	0.77
	Administration & support	Quaternary	0.04	0.04	0.02	1.00	0.53	0.98	0.79	0.95
	Public services	Quaternary	**0.00**	**0.00**	**0.00**	**0.00**	0.05	**0.00**	0.26	0.67
	Other services	Quaternary	0.38	0.86	0.83	**0.00**	1.00	0.04	0.17	0.97

than the Business Survey. Except for the Public services industry that has incremental predictive power for the majority of macro-economic growth indicators, the predictive power of bank sentiment for predicting future macro-economic developments is limited. This is in line with Dell'Ariccia et al. [2008] who find that the real effects of a banking crisis are limited in developed countries, in countries that have more access to foreign financing, and countries where banking crises are less severe, which all apply to Germany.

3.7 Forecasting German macro-economic developments

We perform a rolling-window forecast exercise using a window of length $S = 30$. For each time window, we estimate the 8 ARX models. We use the same selection

Table 3.7: $100 \cdot$ *MAFE for the three selection techniques (rows), the five estimation techniques (columns), and the 8 macro-economic indicators (blocks).*

Selection technique	Response	Estimation technique					Response	Estimation technique				
		Adaptive Lasso	Bayesian	Factor Model	Bagging	Random Forest		Adaptive Lasso	Bayesian	Factor Model	Bagging	Random Forest
All	IP-A1	1.29	0.85	1.19	0.97	0.85	IP-CaGo	2.35	1.63	2.74	1.68	1.47
Granger Lasso Selection		1.23	0.88	1.26	0.91	0.90		2.37	1.63	2.74	1.70	1.57
Granger Lasso test		1.03	0.85	0.88	0.89	0.82		2.28	1.60	2.30	1.67	1.40
All	IP-A2	1.32	0.76	1.13	0.80	0.68	IP-CoGo	1.10	0.60	0.91	1.19	1.12
Granger Lasso Selection		1.29	0.73	1.09	0.83	0.71		1.21	0.61	1.03	1.17	1.12
Granger Lasso test		0.86	0.76	1.20	0.73	0.67		0.79	0.66	0.90	1.15	0.87
All	IP-M	1.51	1.02	1.51	0.99	0.87	RT	1.97	1.07	1.67	0.95	1.00
Granger Lasso Selection		1.60	1.02	1.51	0.98	0.89		1.98	1.06	1.73	0.91	0.97
Granger Lasso test		1.36	1.00	1.29	0.87	0.86		0.98	1.00	1.51	0.96	0.94
All	IP-E	2.46	1.26	2.20	1.03	1.04	WS	1.52	0.52	0.92	0.73	0.63
Granger Lasso Selection		2.47	1.26	2.19	1.28	1.07		1.57	0.52	0.97	0.77	0.63
Granger Lasso test		1.98	1.24	2.21	1.23	1.16		0.79	0.56	0.53	0.64	0.66

and estimation techniques as in Section 3.5.2, except for the standard Wald test and the ML estimator which are not available since the number of time series exceeds the time series length. Next, one-step-ahead forecasts are computed for $t = S + 1, \ldots, T$. We report the Mean Absolute Forecast Error, see equation (3.8), for each macro-economic indicator and each selection-estimation technique combination in Table 3.7.

Among the three selection techniques, the proposed Granger Lasso test performs best. It attains the lowest value of the MAFE in 31 out of 40 cases (78% of the cases). A paired t-test on the 40 MAFEs indicates that the Granger Lasso test significantly outperforms the other selection techniques (both $p-$values< 0.01). Using all industries or using Granger Lasso Selection yields MAFEs close to each other. It turns out that the latter hardly discards industry blocks. In contrast, a much more parsimonious model is obtained using the Granger Lasso test. These parsimonious models lead to an improved forecast accuracy, in the majority of cases.

For the Adaptive Lasso estimation technique, selection based on the Granger Lasso test consistently leads to the lowest MAFE. The MAFEs with the Granger Lasso test are, on average, 21% lower compared to the other selection techniques. After the first selection step where either an entire block of business or bank sentiment indicators is selected or not, the Adaptive Lasso allows some of the time series belonging to one of the selected blocks to be discarded in this second stage. Further reducing the number of relevant predictor time series within the selected blocks improves forecast accuracy.

In line with the results of our simulation study, pre-selecting based on the

Granger Lasso test is less favorable for the Bayesian shrinkage estimator compared to the other estimation techniques. Nevertheless, the Granger Lasso test in combination with the Bayesian shrinkage estimator still leads to the lowest MAFE for 5 out of 8 macro-economic indicators, with an average reduction in MAFE of 5%.

For the Factor Model, the Granger Lasso test leads to the lowest MAFE for 6 out of 8 macro-economic indicators. The MAFEs with the Granger Lasso test are, on average, 20% lower compared to the other selection techniques. Discarding the least predictive industry blocks in this high-dimensional data set and estimating the factors based on the most predictive industry blocks thus leads to important gains in forecast accuracy. This result is in line with Bai and Ng [2008] who find important gains in forecast accuracy from diffusion index models by not using all predictors but by using fewer, informative predictors.

Also for Bagging and Random Forest, the Granger Lasso test selection technique leads towards the lowest MAFE for the majority (6 out of 8) of macro-economic indicators, with an average reduction of 8%. A similar conclusion is drawn by Inoue and Kilian [2008] who find Bagging in combination with pre-selecting predictors to improve macro-economic forecast performance.

3.8 Alternative approaches

We consider three alternative approaches to select the most predictive sentiment indicators and investigate whether they change our findings from Section 3.6 and Section 3.7.[5]

3.8.1 Block size

The proposed Granger Lasso test investigates the predictive power of *blocks* of sentiment indicators. Either all 9 business sentiment indicators are selected for an industry, are none. Similarly, either all 6 bank sentiment indicators are selected for an industry, are none. An advantage of this block approach is that it is decisive: an industry segment is either found to be predictive or not, which eases interpretation. Alternatively, we take blocks of size one and perform $k = 150$ Granger Causality tests, each time testing whether an individual sentiment change indicator (one out of 150) Granger Causes the macro-economic growth indicator.

[5] Detailed results are available from the authors upon request.

We find no significance difference in forecast performance between the Granger Lasso test applied on the blocks, as discussed in Section 3.7, or on the blocks of size one (p-value = 0.22 of paired t-test). It turns out that cost-related assessment questions (i.e. assessment of changes in investments, cost through government red tape, utility and maintenance costs, employees) contain most incremental predictive power. Income-related assessment questions (i.e. assessment of changes in turnover) contain, overall, less incremental predictive power.

3.8.2 Aggregated sentiment indicators

Instead of working with 20 blocks of individual sentiment indicators, we replace them by their average value. This results in a total of 20 aggregated sentiment indicators. We test whether an aggregated sentiment change indicator (one out of 20) Granger Causes the macro-economic growth indicator.

We find no significant difference in forecast performance between the Granger Lasso test applied on the blocks, as discussed in Section 3.7, or on the aggregated sentiment indicators (p-value=0.90 of paired t-test). In line with our previous results, we find that (i) the Business Survey contains more incremental predictive power than the Bank Survey, (ii) industries contain most predictive power for those macro-econonmic indicators most closely tied to their day-to-day business.

3.8.3 Segmentation criterion

Our main research question is whether the sentiment of different industry segments has predictive power for macro-economic indicators. Our methodology is also applicable to other ways of segmenting firms, as *region* in which the are located or according to their *company size*. For our data, there are 10 regions and three company sizes. We re-estimate the 8 ARX models and perform the Granger Causality tests for the 20 regional blocks (i.e. 10 blocks for the Business Survey, 10 blocks for the Bank Survey). Likewise, we re-estimate the 8 ARX models and perform the Granger Causality tests for the 6 company size blocks (i.e. 3 blocks for the Business Survey, 3 blocks for the Bank Survey).

The forecast performance of the Granger Lasso test obtained with either industry, region or company size segments is very similar. We compare Mean Absolute Forecast Errors as in Table 3.7. For the regional segments, the Granger Lasso test is the best performing selection technique and attains the lowest value of the MAFE in 60% of the cases (24 out of 40). Similarly for the company size segments

where the Granger Lasso test leads towards the lowest MAFE in 68% of the cases (27 out of 40).

We again find business sentiment to have more incremental predictive power than bank sentiment. Furthermore, Germany's largest geo-economical regions (i.e. Ruhr area and the Southern states) have most incremental predictive power for the macro-economic indicators to which their day-to-day business contributes most, i.e. IP-A1, IP-A2, IP-M, IP-E and IP-CaGo, IP-CoGo respectively. Finally, small- and medium-sized companies have more incremental predictive power than large companies. Germany is dominated by small- to medium-sized companies who are global market leaders in their segments, and, hence, those might be best at evaluating Germany's economy.

3.9 Discussion

This paper presents a high-dimensional Granger Causality test. It detects the most predictive industry segments for future macro-economic developments. For this purpose, we use both business and bank sentiment surveys answered by firms across Germany. Not all industry-specific sentiment indicators are equally predictive for all macro-economic indicators. Industries contain most predictive power for the macro-economic indicators most closely tied to their day-to-day business activities.

Our forecast exercise shows that important gains in forecast accuracy can be obtained by not using all industry segments, but by first selecting the most predictive ones using the Granger Lasso test selection technique. In high-dimensional settings, a lot of noise might be present. By selecting predictor variables, a more parsimonious model with less noise is obtained. Note that losing information is a potential risk of selecting predictor variables, hence, the need for research on appropriate selection methods. The selection of the most pertinent industry segments also provides important information for institutes conducting these sentiment surveys. For instance, instead of equally spreading respondents among all segments, the number of respondents in predictive segments could be increased, whereas the number of respondents in non-predictive segments could be decreased. Alternatively, non-predictive segments could even be completely discarded, which provides an opportunity to obtain cost savings.

The identification of pertinent respondents also applies to consumer sentiment surveys. In the large literature on consumer sentiment, this topic has received little attention. We perform a similar exercise as described in this paper using a

consumer sentiment survey data set from the National Bank of Belgium. Sentiment indicators are available for different classes of consumers' net disposable income, profession, employment status, education, age and gender. We study their predictive power for several retail trade indicators. The profession, education, and age sentiment indicators contain most predictive power. Again, important gains in forecast accuracy can be obtained by first selecting the most predictive sentiment indicators (for a specific target variable of interest) instead of using all indicators.

We use a high-dimensional Granger Causality approach to study the predictive power of sentiment data collected via surveys. One could consider social media as an alternative channel to collect sentiment data. While their role in collecting consumer sentiment has received considerable attention (e.g. Pang and Lee, 2008, Asur and Huberman, 2010, Stieglitz and Dang-Xuan, 2013), their role in collecting business sentiment has received limited to no attention. It should be noted that collecting data via social media poses sampling issues since only a subpopulation (i.e. the participants of these social media) of all respondents is reached. In contrast, data collected via surveys are sent out to a random sample of all respondents.

While we study the predictive power of sentiment indicators for future macro-economic growth, another interesting research question is whether sentiment indicators and macro-economic indicators move together in the long-run. This could be addressed using Cointegration analysis, which aims at detecting long-run relationships between several time series (see Lütkepohl, 1993 for an introduction, Östermark, 2001 or Musti and D'Ecclesia, 2008 for an application. Testing for cointegration in high-dimensions is, however, an open research area (e.g. Breitung and Cubadda, 2011) and ideas similar to the once introduced in this paper could serve as a starting point.

Finally, we need to further deepen our understanding on the usefulness of bank sentiment. It would be interesting to investigate if this sentiment differs between, for instance, countries that are more or less severely hit by banking crises, and developed or developing countries. The study of sentiment with respect to the banking sector opens a new area of research on sentiment surveys.

Chapter 4

Forecasting using sparse cointegration

Abstract

This paper proposes a sparse cointegration method. Cointegration analysis is used to estimate the long-run equilibrium relations between several time series. The coefficients of these long-run equilibrium relations are the cointegrating vectors. We provide a sparse estimator of the cointegrating vectors. Sparse estimation means that some elements of the cointegrating vectors are estimated as exactly zero, improving interpretability. The sparse estimator is applicable in high-dimensional settings, where the length of the time series is short compared to the number of time series. Our method achieves better estimation accuracy and forecast accuracy than the traditional Johansen method in sparse and/or high-dimensional settings. We use the sparse method for interest rate growth forecasting and consumption growth forecasting. The sparse cointegration method leads to important gains in forecast accuracy compared to the Johansen method.

4.1 Introduction

High-dimensional data sets containing thousands of time series are commonly available and accessible at reasonable cost [Stock and Watson, 2002, Fan et al., 2011]. There has been a considerable amount of recent work exploiting the large amount of information in these data sets for forecasting purposes. To handle the

dimensionality, large time series models, containing a large number of time series relative to the time series length, have been considered. Common approaches are, among others, Factor Models (e.g. Stock and Watson, 2002), Bayesian Vector AutoRegressive (VAR) Models (e.g. Banbura et al., 2010), or Reduced-Rank VAR Models (e.g. Carriero et al., 2011, Bernardini and Cubadda, 2015). Typically, these authors do not account for cointegration. Instead, the time series are either transformed in order to achieve stationarity [Bernardini and Cubadda, 2015] or the (non)-stationarity is accounted for in the prior distribution of the autoregressive parameters [Banbura et al., 2010]. In cointegration analysis, long-run equilibrium relations between several time series, often implied by economic theory, are estimated.

This paper develops a cointegration method for high-dimensional time series. The Vector Error Correcting Model (VECM) (e.g. Lütkepohl, 2007) is used to estimate and test for the cointegration relations. Various cointegration tests are existing (e.g. Engle and Granger, 1987, Phillips and Ouliaris, 1990), among which the cointegration test of Johansen [1988] has become most popular. Johansen's Maximum Likelihood approach has, however, some limitations. In high-dimensional settings, where the number of time series is large compared to the length of the time series, the estimation imprecision will be large. Johansen's approach is based on the estimation of a VAR model and a canonical correlation analysis. A drawback of the VAR is that its number of parameters increases quadratically with the number of included time series. Consequently, regression parameters are estimated inaccurately if only a limited number of time points is available. When the number of time series exceeds the time series length, Johansen's approach can not even be applied.

We introduce a Penalized Maximum Likelihood (PML) approach to estimate the cointegrating vectors in a sparse way, i.e. some of its components are estimated as exactly zero. Sparse estimators show good performance in various fields such as economics (e.g. Fan et al., 2011), macro-economics (e.g. Korobilis, 2013; Liao and Phillips, 2015), finance (e.g. Zhou et al., 2014), or biostatistics (e.g. Friedman, 2012). A sparse cointegration method is useful for several reasons. First, sparsity facilitates model interpretation since only a limited number of time series, those corresponding to the non-zero coefficients, enter the estimated long-run equilibrium relations. Second, sparsity improves forecast performance through variance reduction. Third, the sparse approach, in contrast to Johansen's Maximum Likelihood approach, can be applied when the number of time series exceeds the time series length.

We show in a simulation study that the sparse cointegration method significantly outperforms Johansen's method when the cointegrating vectors are sparse or when the number of time series is large compared to the time series length. Furthermore, we evaluate the forecast performance of the proposed sparse cointegration method on two data sets. We show that important gains in forecast accuracy can be obtained by accounting for cointegration and by sparsely estimating the cointegrating vectors.

The remainder of this article is structured as follows. We describe the sparse cointegration method in Section 4.2. Section 4.3 provides more details on the algorithm. Section 4.4 discusses the Rank Selection Criterion [Bunea et al., 2011] to determine the cointegration rank. Section 4.5 presents the results of a simulation study. Section 4.6 discusses two forecasting examples. First we forecast interest rate growth, secondly we forecast consumption growth. Finally, Section 4.7 concludes.

4.2 Penalized Maximum Likelihood

Let $\mathbf{y}_t$ be a q-dimensional multivariate time series. We assume that the vector process $\mathbf{y}_t$ is integrated of order one $I(1)$, meaning that its first difference is stationary. Note that $\mathbf{y}_t$ can be $I(1)$ even though some of its components are stationary (Chapter 5 Johansen, 1991). Furthermore, we assume that $\mathbf{y}_t$ follows a Vector AutoRegressive model of order p, denoted as VAR(p). Any p^{th} order VAR can be re-written in Vector Error Correcting (VECM) representation [Hamilton, 1991] as follows

$$\boldsymbol{\Delta}\mathbf{y}_t = \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \boldsymbol{\Delta}\mathbf{y}_{t-i} + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad t = p+1, \ldots, T \tag{4.1}$$

where $\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_{p-1}$ are $q \times q$ matrices containing short-run effects, $\boldsymbol{\Pi}$ is a $q \times q$ matrix of rank r, $0 \leq r \leq q$ and $\boldsymbol{\varepsilon}_t$ is assumed to follow a $N_q(\mathbf{0}, \boldsymbol{\Sigma})$.

If we can express $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T$ with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ $q \times r$ matrices of full column rank r, with $0 < r < q$, then the linear combinations given by $\boldsymbol{\beta}^T\mathbf{y}_t$ are stationary and $\mathbf{y}_t$ is said to be cointegrated with cointegration rank r. The cointegrating vectors are the columns of $\boldsymbol{\beta}$ and the adjustment coefficients the elements of $\boldsymbol{\alpha}$.

We estimate the model parameters by Penalized Maximum Likelihood (PML). It is convenient to rewrite model (4.1) in matrix notation:

$$\boldsymbol{\Delta}\mathbf{Y} = \boldsymbol{\Delta}\mathbf{Y}_\mathbf{L}\boldsymbol{\Gamma} + \mathbf{Y}\boldsymbol{\Pi}^T + \mathbf{E} \tag{4.2}$$

where $\mathbf{\Delta Y} = (\mathbf{\Delta y}_{p+1}, \ldots, \mathbf{\Delta y}_T)^T$; $\mathbf{\Delta Y_L} = (\mathbf{\Delta X}_{p+1}, \ldots, \mathbf{\Delta X}_T)^T$ with $\mathbf{\Delta X}_t = (\mathbf{\Delta y}_{t-1}^T, \ldots, \mathbf{\Delta y}_{t-p+1}^T)^T$; $\mathbf{Y} = (\mathbf{y}_p, \ldots, \mathbf{y}_{T-1})^T$; $\mathbf{\Gamma} = (\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_{p-1})^T$; and $\mathbf{E} = (\boldsymbol{\varepsilon}_{p+1}, \ldots, \boldsymbol{\varepsilon}_T)^T$.

Consider the penalized negative log-likelihood

$$\mathcal{L}(\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}) = \frac{1}{T}\mathrm{tr}\Big((\mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma} - \mathbf{Y}\mathbf{\Pi}^T)\mathbf{\Omega}(\mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma} - \mathbf{Y}\mathbf{\Pi}^T)^T\Big) - \log|\mathbf{\Omega}| + \\ + \lambda_1 P_1(\boldsymbol{\beta}) + \lambda_2 P_2(\mathbf{\Gamma}) + \lambda_3 P_3(\mathbf{\Omega}), \quad (4.3)$$

with $\mathrm{tr}(\cdot)$ denoting the trace, $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, and P_1, P_2 and P_3 three penalty functions.

We use L_1 penalization (see the Lasso estimator, Tibshirani, 1996) on the cointegrating vectors $\boldsymbol{\beta}$

$$P_1(\boldsymbol{\beta}) = \sum_{i=1}^{q}\sum_{j=1}^{r}|\beta_{ij}|. \quad (4.4)$$

By adding the L_1 penalty to the objective function in (4.3), a sparse solution is obtained: some elements of $\boldsymbol{\beta}$ are estimated as exactly zero. Similarly, we use L_1 penalization on the short-run effects $\mathbf{\Gamma}$ and the off-diagonal elements of the inverse of the error covariance matrix $\mathbf{\Omega}$.

The aim is to select $\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}$ so as to minimize (4.3) subject to the constraint

$$\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T,$$

with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ $q \times r$ matrices of full column rank r. The matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not uniquely defined. For identifiability purposes, we impose the normalization conditions $\boldsymbol{\alpha}^T\mathbf{\Omega}\boldsymbol{\alpha} = \mathbf{I}_r$. For the unpenalized case ($\lambda_1 = 0$, $\lambda_2 = 0$ and $\lambda_3 = 0$), the objective function (4.3) boils down to the one introduced by Johansen [1988]. The unpenalized case can be solved by the closed-form expressions documented in Johansen [1988] or by using the iterative algorithm described below.

4.3 Algorithm

To find the minimum of the penalized negative log-likelihood in (4.3), we iteratively solve for $\mathbf{\Pi}$ conditional on $\mathbf{\Gamma}, \mathbf{\Omega}$; for $\mathbf{\Gamma}$ conditional on $\mathbf{\Pi}, \mathbf{\Omega}$; and for $\mathbf{\Omega}$ conditional on $\mathbf{\Gamma}, \mathbf{\Pi}$.

Solving for $\mathbf{\Pi}$ conditional on $\mathbf{\Gamma}, \mathbf{\Omega}$. When $\mathbf{\Gamma}$ and $\mathbf{\Omega}$ are fixed, the minimization

problem in (4.3) with $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T$ is equivalent to

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})|\mathbf{\Gamma}, \mathbf{\Omega} = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{T}\operatorname{tr}\Big((\mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^T)\mathbf{\Omega}(\mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^T)^T\Big) + \\ + \lambda_1 P_1(\boldsymbol{\beta}), \quad (4.5)$$

which boils down to a penalized reduced rank regression [Chen et al., 2012]. We first estimate $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$, next we estimate $\boldsymbol{\beta}$ conditional on $\boldsymbol{\alpha}$.

For fixed $\boldsymbol{\beta}$, the minimization problem in (4.5) reduces to

$$\hat{\boldsymbol{\alpha}}|\mathbf{\Gamma}, \mathbf{\Omega}, \boldsymbol{\beta} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{T}\operatorname{tr}\Big((\mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^T)\mathbf{\Omega}(\mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^T)^T\Big),$$

subject to $\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{\Omega}\boldsymbol{\alpha} = \mathbf{I}_r$, which is a weighted Procrustes problem [Lissitz et al., 1976]. This weighted Procrustes problem for $\boldsymbol{\alpha}$ can be seen as an unweighted Procrustes problem for $\boldsymbol{\alpha}^\star = \mathbf{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}$. The solution is

$$\hat{\boldsymbol{\alpha}} = \mathbf{\Omega}^{-1/2}\mathbf{V}\mathbf{U}^{\mathbf{T}},$$

where $\mathbf{U}$ and $\mathbf{V}$ are obtained from the singular value decomposition of

$$\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{Y}^T(\mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma})\mathbf{\Omega}^{\mathbf{1/2}} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}}.$$

Chen et al. [2012] only consider the case where $\mathbf{\Omega} = \mathbf{I}$, and use a Procrustes problem to solve for $\boldsymbol{\alpha}$. A weighted Procrustes problem takes the covariance structure into account.

For fixed $\boldsymbol{\alpha}$, the minimization problem in (4.5) reduces to

$$\hat{\boldsymbol{\beta}}|\mathbf{\Gamma}, \mathbf{\Omega}, \boldsymbol{\alpha} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{T}\operatorname{tr}\Big((\mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^T)\mathbf{\Omega}(\mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^T)^T\Big) + \\ + \lambda_1 P_1(\boldsymbol{\beta}). \quad (4.6)$$

Since $\boldsymbol{\alpha}^{\boldsymbol{\star T}}\boldsymbol{\alpha}^\star = \mathbf{I}_r$, there exists a matrix $\boldsymbol{\alpha}^{\star\perp}$ with orthonormal columns such that $(\boldsymbol{\alpha}^\star, \boldsymbol{\alpha}^{\star\perp})$ is an orthogonal matrix. Then, with $\tilde{\mathbf{Y}} = \mathbf{\Delta Y} - \mathbf{\Delta Y_L}\mathbf{\Gamma}$,

$$\begin{aligned}
\operatorname{tr}\Big((\tilde{\mathbf{Y}} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^T)\mathbf{\Omega}(\tilde{\mathbf{Y}} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^T)^T\Big) &= ||(\tilde{\mathbf{Y}} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^T)\mathbf{\Omega}^{\mathbf{1/2}}||^2 \\
&= ||(\tilde{\mathbf{Y}}\mathbf{\Omega}^{\mathbf{1/2}} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^{\star T})||^2 \\
&= ||(\tilde{\mathbf{Y}}\mathbf{\Omega}^{\mathbf{1/2}} - \mathbf{Y}\boldsymbol{\beta}\boldsymbol{\alpha}^{\star T})(\boldsymbol{\alpha}^\star, \boldsymbol{\alpha}^{\star\perp})||^2 \\
&= ||\tilde{\mathbf{Y}}\mathbf{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^\star - \mathbf{Y}\boldsymbol{\beta}||^2 + ||\tilde{\mathbf{Y}}\mathbf{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^{\star\perp}||^2,
\end{aligned}$$

where $||\cdot||$ denotes the Frobenius norm for a matrix. Since the second term on the left-hand-side does not involve $\boldsymbol{\beta}$, the minimization problem reduces to

$$\hat{\boldsymbol{\beta}}|\boldsymbol{\Gamma},\boldsymbol{\Omega},\boldsymbol{\alpha} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad \frac{1}{T}\operatorname{tr}\Big((\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^{\star} - \mathbf{Y}\boldsymbol{\beta})(\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^{\star} - \mathbf{Y}\boldsymbol{\beta})^T\Big) + \lambda_1 P_1(\boldsymbol{\beta}), \tag{4.7}$$

which is a penalized multivariate least squares regression of $\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^{\star}$ on $\mathbf{Y}$.

Solving for $\boldsymbol{\Gamma}$ conditional on $\boldsymbol{\Pi}, \boldsymbol{\Omega}$. When $\boldsymbol{\Pi}$ and $\boldsymbol{\Omega}$ are fixed, the minimization problem in (4.3) is a penalized multivariate regression of $\big(\boldsymbol{\Delta}\mathbf{Y} - \mathbf{Y}\boldsymbol{\Pi}^T\big)$ on $\boldsymbol{\Delta}\mathbf{Y}_{\mathbf{L}}$, see Rothman et al. [2010].

Solving for $\boldsymbol{\Omega}$ conditional on $\boldsymbol{\Gamma}, \boldsymbol{\Pi}$. When $\boldsymbol{\Gamma}$ and $\boldsymbol{\Pi}$ are fixed, the minimization problem in (4.3) corresponds to penalized covariance estimation [Friedman et al., 2008].

Convergence criterion. We iterate solving the minimization problems described above until the relative change in the value of the objective function, i.e. the penalized log-likelihood in (4.3), in two successive iterations[1] is smaller than a prespecified tolerance level ϵ, chosen to be $\epsilon = 10^{-2}$. Although there is no proof of convergence of the algorithm, we have observed it empirically in all real data examples and all simulation runs. For a data set (generated as in the Simulation Study of Section 5) consisting of $q = 4$ time series each of length $T = 500$, on average three iterations were needed to converge, for $q = 11$ and $T = 50$, on average four iterations were needed to converge.

Selection of tuning parameters. Tuning parameters are selected in each step of the iterative algorithm. We select the tuning parameters λ_1, controlling the penalization on the cointegrating vectors, and λ_2, controlling the penalization of the short-run effects, according to a time series cross-validation approach [Hyndman, 2014], see Appendix 4.8. The tuning parameter λ_3, controlling the penalization on the off-diagonal elements of $\boldsymbol{\Omega}$, is selected according to the Bayesian Information Criterion [Friedman et al., 2008]. As a default, we use a grid of hundred λ_1 values, five λ_2 values and five λ_3 values.

Starting values. A starting value for $\boldsymbol{\Omega}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\beta}$ is required. We take the identity matrices for $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}_k, k = 1, \ldots, p-1$. For $\boldsymbol{\beta}$ we take the first r eigenvectors of

[1] One iteration includes one cycle of estimating $\boldsymbol{\Pi}|\boldsymbol{\Gamma},\boldsymbol{\Omega}$; $\boldsymbol{\Gamma}|\boldsymbol{\Pi},\boldsymbol{\Omega}$; and $\boldsymbol{\Omega}|\boldsymbol{\Gamma},\boldsymbol{\Pi}$.

the matrix $\widehat{\Sigma}_{YY}^{-1}\widehat{\Sigma}_{Y\Delta Y}\hat{\Sigma}_{\Delta Y\Delta Y}^{-1}\widehat{\Sigma}_{\Delta YY}$, where we take $\widehat{\Sigma}_{YY}$ and $\widehat{\Sigma}_{\Delta Y\Delta Y}$ diagonal and $\widehat{\Sigma}_{Y\Delta Y} = \widehat{\Sigma}_{\Delta YY}^{T}$ the sample covariance matrix between $\mathbf{Y}$ and $\mathbf{\Delta Y}$.

We performed several numerical experiments to investigate the robustness of the outcome of the algorithm to the choice of starting values. In low-dimensional settings, the choice of starting values is not important. In high-dimensional settings, a good choice of starting values is more important. Note that the starting values should exist and be easily computable in all settings, which holds for our proposal.

Computation time. All computations are carried out in `R` version 3.2.1. The PML estimator is rather fast to compute. Computation time includes the cross-validation for the selection of tuning parameters. For a data set consisting of $q = 4$ time series each of length $T = 500$, on average 8 seconds are needed on an Intel Core i7-3720QM @ 2.60GHz machine. For a data set with $q = 11$ and $T = 50$, we need 4 seconds on average.

4.4 Determination of Cointegration Rank

At small finite samples, the asymptotic distribution of Johansen's trace statistic, used to determine the cointegration rank, might poorly approximate the true distribution, resulting in substantial size and power distortions (e.g. Johansen, 2002; Nielsen, 2004; Breitung and Cubadda, 2011). We use an iterative procedure based on the Rank Selection Criterion (RSC) of Bunea et al. [2011] to determine the cointegration rank r. We start with an initial value of the cointegration rank $r_{\text{start}} = q$.

For this initial value, we obtain $\widehat{\mathbf{\Gamma}}$ using the algorithm from Section 4.3. Next, we update our estimate of the cointegration rank. Following Bunea et al. [2011], $\hat{r}$ is given by the number of eigenvalues of the matrix $\widetilde{\mathbf{\Delta Y}}^{T}\mathbf{P}\widetilde{\mathbf{\Delta Y}}$ that exceeds the threshold μ:

$$\hat{r} = \max\{r : \lambda_r(\widetilde{\mathbf{\Delta Y}}^{T}\mathbf{P}\widetilde{\mathbf{\Delta Y}}) \geq \mu\},$$

with $\widetilde{\mathbf{\Delta Y}} = \mathbf{\Delta Y} - \mathbf{\Delta Y_L}\widehat{\mathbf{\Gamma}}$ and $\mathbf{P} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-}\mathbf{Y}^T$ the projection matrix onto the column space of $\mathbf{Y}$. Note that $(\mathbf{Y}'\mathbf{Y})^{-}$ denotes the Moore-Penrose inverse of the matrix $(\mathbf{Y}'\mathbf{Y})$. Following the recommendation of Bunea et al. [2011], the threshold is set equal to $\mu = 2S^2(q + l)$, under the assumption that $l < T$, with $l = \text{rank}(\mathbf{Y})$ and

$$S^2 = \frac{||\widetilde{\mathbf{\Delta Y}} - \mathbf{P}\widetilde{\mathbf{\Delta Y}}||^2}{Tq - lq}.$$

We repeat the above procedure using the new value of $\hat{r}$, this until the estimated cointegration rank does not change in two successive iterations.

The Rank Selection Criterion consistently estimates the effective rank of the coefficient matrix $\mathbf{\Pi}$ in the penalized reduced rank regression [Bunea et al., 2011]. The consistency results are valid when either the length of the time series or the number of time series grows to infinity. This procedure to determine the rank has almost no computational cost.

4.5 Simulation Studies

We conduct a simulation study to evaluate the performance of the PML estimator. The data generating process (revised from Cavaliere et al., 2012) is the following VECM:

$$\mathbf{\Delta y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}^T\mathbf{y}_{t-1} + \mathbf{\Gamma}_1\mathbf{\Delta y}_{t-1} + \mathbf{e}_t, \qquad (t = p+1, \ldots, T),$$

where the error terms $\mathbf{e}_t$ follow a $N_q(\mathbf{0}, \mathbf{I}_q)$ distribution. We set $\mathbf{y}_0 = \mathbf{\Delta y}_0 = \mathbf{0}$. All simulated models satisfy the assumptions of the VECM described in Section 4.2.

We compare the out-of-sample forecast accuracy of the PML estimator to the ML estimator of Johansen [1988] and find that the former obtains a significant better forecast performance than the latter in sparse and/or high-dimensional settings. Besides, we also compare their estimation accuracy and investigate the performance of the Rank Selection Criterion in correctly selecting the true cointegration rank.

4.5.1 Simulation designs

Different simulation designs are considered: (i) low-dimensional ($T = 500, q = 4$), and (ii) high-dimensional with moderate time series length ($T = 50, q = 11$)[2]. We consider sparse and non-sparse settings and report on selected representative cases below. Full details on each selected setting are in Table 4.1.

Low-dimensional designs. The true cointegrating vectors and the short-run effects are sparse in the first two simulation settings, non-sparse in the third setting. The cointegration rank equals $r = 1$, $r = 2$, $r = 1$ respectively. While $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ belong to a different space in the first and third setting, they belong to

[2] $q = 11$ time series is the largest number for which the critical values of Johansen's trace statistic are tabulated in Johansen (Chapter 15; 1996).

Table 4.1: *Low-dimensional ($T = 500, q = 4$) and high-dimensional ($T = 50, q = 11$) simulation designs.*

Low-dimensional designs	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\Gamma}_1$	$\boldsymbol{\Sigma}$
Sparse $r = 1$	$\begin{bmatrix} 1 \\ \mathbf{0}_{3\times1} \end{bmatrix}$	$a \cdot \begin{bmatrix} \mathbf{1}_{2\times1} \\ \mathbf{0}_{2\times1} \end{bmatrix}$	$\gamma \boldsymbol{I}_q$	$\mathbf{I}_q$
Sparse $r = 2$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} \end{bmatrix}$	$a\boldsymbol{\beta}$	$\gamma \boldsymbol{I}_q$	$\Sigma_{ij} = 0.2^{\lvert i-j \rvert}$
Non-sparse $r = 1$	$\begin{bmatrix} 1 \\ \mathbf{0.1}_{3\times1} \end{bmatrix}$	$a \cdot \begin{bmatrix} \mathbf{1}_{2\times1} \\ \mathbf{0.1}_{2\times1} \end{bmatrix}$	$\boldsymbol{\Gamma}_{1,ij} = \begin{cases} \gamma & if\ j = i \\ \gamma \cdot 10^{-4} & if\ j \neq i \end{cases}$	$\gamma \boldsymbol{I}_q$
	with $a = -0.2, -0.4, \ldots, -0.8$, and $\gamma = 0.1$			
High-dimensional designs	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\Gamma}_1$	$\boldsymbol{\Sigma}$
Sparse $r = 1$	$\begin{bmatrix} \mathbf{1}_{3\times1} \\ \mathbf{0}_{8\times1} \end{bmatrix}$	$a \cdot \begin{bmatrix} \mathbf{1}_{6\times1} \\ \mathbf{0}_{5\times1} \end{bmatrix}$	$\gamma \boldsymbol{I}_q$	$\mathbf{I}_q$
Sparse $r = 4$	$\begin{bmatrix} \mathbf{1}_{3\times1} & \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} \\ \mathbf{0}_{3\times1} & \mathbf{1}_{3\times1} & \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} \\ \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} & \mathbf{1}_{3\times1} & \mathbf{0}_{3\times1} \\ \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} & \mathbf{1}_{2\times1} \end{bmatrix}$	$a\boldsymbol{\beta}$	$\gamma \boldsymbol{I}_q$	$\Sigma_{ij} = 0.2^{\lvert i-j \rvert}$
Non-sparse $r = 1$	$\begin{bmatrix} \mathbf{1}_{3\times1} \\ \mathbf{0.1}_{8\times1} \end{bmatrix}$	$a \cdot \begin{bmatrix} \mathbf{1}_{6\times1} \\ \mathbf{0.1}_{5\times1} \end{bmatrix}$	$\boldsymbol{\Gamma}_{1,ij} = \begin{cases} \gamma & if\ j = i \\ \gamma \cdot 10^{-4} & if\ j \neq i \end{cases}$	$\gamma \boldsymbol{I}_q$
	with $a = -0.2, -0.4, \ldots, -0.8$ and $\gamma = 0.4$			

the same space in setting two. Furthermore, the error terms of the VECM are uncorrelated in setting one and three, they are correlated in the second setting.

High-dimensional designs. The true cointegrating vectors and the short-run effects are sparse in the first two simulation settings, non-sparse in the third setting. The cointegration rank equals $r = 1$, $r = 4$, $r = 1$ respectively. Choices for the relation between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and the error terms are similar to the low-dimensional designs.

4.5.2 Estimation accuracy

To evaluate the estimation accuracy, we compute for each simulation run m, with $m = 1, \ldots, M = 500$, the angle $\theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta})$ between the estimated cointegration space and the true cointegration space.[3] The average angle is then given by

$$\theta(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{M} \sum_{m=1}^{M} \theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta}). \tag{4.8}$$

[3] The angle $\theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta})$ is computed as follows (see e.g. Anderson, 1958). We compute the QR-decompositions $\widehat{\boldsymbol{\beta}}^m = \mathbf{Q}_{\widehat{\boldsymbol{\beta}}^m} \mathbf{R}_{\widehat{\boldsymbol{\beta}}^m}$ and $\boldsymbol{\beta} = \mathbf{Q}_{\boldsymbol{\beta}} \mathbf{R}_{\boldsymbol{\beta}}$. Next, compute the singular value decomposition of $\mathbf{Q}^T_{\widehat{\boldsymbol{\beta}}^m} \mathbf{Q}_{\boldsymbol{\beta}} = \mathbf{U}\mathbf{C}\mathbf{V}^{\mathbf{T}}$. The matrix $\mathbf{C}$ is diagonal with elements $c_1 \geq \ldots \geq c_r$. The minimum angle is given by $\theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta}) = \cos^{-1}(c_1)$.

Table 4.2: *Average angle between the estimated and true cointegration space. The results are reported for different values of the adjustment coefficient a and dimension q of the VECM. Significant differences, at the 5% significance level, between the PML and ML estimator are in bold.*

Method \ a	-0.2	-0.4	-0.6	-0.8	-0.2	-0.4	-0.6	-0.8
	Low-dimensional				**High-dimensional**			
	Sparse $q = 4, T = 500, r = 1$				Sparse $q = 11, T = 50, r = 1$			
ML	0.032	0.016	0.011	0.008	1.044	0.796	0.559	0.409
PML	**0.020**	**0.010**	**0.007**	**0.005**	**0.588**	**0.226**	**0.160**	**0.138**
	Sparse $q = 4, T = 500, r = 2$				Sparse $q = 11, T = 50, r = 4$			
ML	0.007	0.004	0.003	0.002	0.167	0.088	0.058	0.043
PML	**0.006**	**0.003**	**0.002**	**0.001**	**0.138**	**0.065**	**0.041**	**0.029**
	Non-sparse $q = 4, T = 500, r = 1$				Non-sparse $q = 11, T = 50, r = 1$			
ML	**0.032**	**0.016**	**0.011**	**0.008**	1.045	0.775	0.542	0.384
PML	0.037	0.019	0.013	0.009	**0.646**	**0.289**	**0.220**	**0.248**

The value of the angle varies from zero (for identical subspaces) to $\pi/2$ (for orthogonal subspaces).

Results. Simulation results on the accuracy of the estimated cointegration space are in Table 4.2. For different values of the adjustment coefficients a, we report the average angle (averaged across simulation runs) between the estimated and the true cointegration space. We use a two-sided paired t-test to test equality of the average angle of the PML and ML estimator.

In the sparse low-dimensional settings, the sparse estimator is the best performing. It provides significantly more precise estimates than Johansen's estimator, for almost all values of the adjustment coefficients. In the non-sparse low-dimensional setting, Johansen's ML estimator is best performing, as expected. The usage of the PML procedure does not lead to a lower estimation precision here.

In the high-dimensional designs, the advantage of the PML estimator becomes much larger. The time series length is short compared to the number of time series, such that the estimation imprecision of Johansen's ML estimator will become large. In all settings, the PML estimator indeed significantly outperforms Johansen's ML estimator. Also for the non-sparse setting the PML estimator performs best. The differences are outspoken. Since the PML estimator performs regularization, its good performance is retained in the non-sparse high-dimensional setting.

4.5.3 Forecast accuracy

To evaluate the out-of-sample forecast accuracy, we use a rolling window with window size S. Let h be the forecast horizon. At each time point $t = S, \ldots, T-h$, we use the PML or Johansen's ML estimator to estimate the VECM

$$\widehat{\Delta \mathbf{y}}_{t+h} = \sum_{i=1}^{p-1} \widehat{\mathbf{\Gamma}}_i \mathbf{\Delta y}_{t+1-i} + \widehat{\mathbf{\Pi}} \mathbf{y}_t, \tag{4.9}$$

for different forecast horizons $h \in \{1, 3, 6, 12\}$. h-step-ahead multivariate forecast errors $\widehat{\mathbf{e}}_{t+h} = \Delta \mathbf{y}_{t+h} - \widehat{\mathbf{\Delta y}}_{t+h}$ are obtained. In each simulation run, the overall multivariate forecast performance is then measured by the Multivariate Mean Absolute Forecast Error (e.g. Carriero et al., 2011):

$$\text{MMAFE} = \frac{1}{T-h-S+1} \sum_{t=S}^{T-h} \frac{1}{q} \sum_{i=1}^{q} \frac{|\Delta y_{t+h}^{(i)} - \widehat{\Delta y}_{t+h}^{(i)}|}{\widehat{\sigma}_{(i)}}, \tag{4.10}$$

where $\widehat{\sigma}_{(i)}$ is the standard deviation of the i^{th} time series in differences. The MMAFE depends on the forecast horizon h.

For the low-dimensional designs, we consider window sizes $S \in \{48, 96, 144\}$. The window size S is the number of time points available for estimation. We expect the gain in forecast performance of the PML estimator relative to the ML estimator to be larger for small values of S. For the high-dimensional designs, we only consider a window size $S = 36$ to have sufficient time points available for the estimation of the models.

Results. Simulation results on the out-of-sample forecast accuracy in the low-dimensional designs are in Table 4.3. For reasons of brevity, we only report the results for $a = -0.4$. The MMAFE is computed for four different forecast horizons (columns Table 4.3) and three rolling window sizes (rows Table 4.3). The PML estimator always attains a lower value of the MMAFE than the ML estimator. A two-sided paired t-test confirms that these improvements in forecast performance are significant (all p-values< 0.01). Also in the non-sparse low-dimensional setting the forecast accuracy of the PML estimator is better than the one of the ML estimator, though the difference between both is small especially for $S = 144$. Regardless of the degree of sparsity of the cointegrating vector (i.e. the number of zero components in the cointegrating vector), the largest gain in forecast accuracy of the PML relative to the ML estimator is attained when the rolling window size is the lowest ($S = 48$), and this for all forecast horizons. Furthermore, the forecast performance of the PML estimator is stable for the

Table 4.3: *Low-dimensional designs. Multivariate Mean Absolute Forecast Error using the PML and ML estimator. For each window size S (rows) - forecast horizon h (columns) combination, the lowest values are indicated in bold.*

Setting	$h=1$		$h=3$		$h=6$		$h=12$	
Window size S	PML	ML	PML	ML	PML	ML	PML	ML
Sparse $q=4, T=500, r=1$								
$S=48$	**0.85**	0.88	**0.84**	0.88	**0.84**	0.89	**0.84**	0.89
$S=96$	**0.84**	0.85	**0.83**	0.84	**0.83**	0.85	**0.83**	0.85
$S=144$	**0.83**	0.84	**0.82**	0.83	**0.82**	0.84	**0.82**	0.84
Sparse $q=4, T=500, r=2$								
$S=48$	**0.88**	0.97	**0.88**	0.95	**0.88**	0.96	**0.88**	0.96
$S=96$	**0.87**	0.93	**0.87**	0.91	**0.87**	0.92	**0.87**	0.92
$S=144$	**0.87**	0.91	**0.87**	0.90	**0.87**	0.90	**0.87**	0.91
Non-sparse $q=4, T=500, r=1$								
$S=48$	**0.86**	0.89	**0.85**	0.88	**0.85**	0.90	**0.85**	0.89
$S=96$	**0.85**	0.86	**0.84**	0.85	**0.83**	0.85	**0.84**	0.86
$S=144$	**0.84**	0.85	**0.83**	0.84	**0.83**	0.84	**0.83**	0.84

Table 4.4: *High-dimensional designs. Multivariate Mean Absolute Forecast Error using the PML and ML estimator. For each forecast horizon h, the lowest values are indicated in bold.*

	$h=1$		$h=3$		$h=6$		$h=12$	
Setting	PML	ML	PML	ML	PML	ML	PML	ML
Sparse $q=11, T=50, r=1$	**0.87**	0.91	**0.88**	1.08	**0.87**	1.07	**0.87**	1.06
Sparse $q=11, T=50, r=4$	**0.98**	1.16	**0.99**	1.28	**0.98**	1.26	**0.97**	1.27
Non-sparse $q=11, T=50, r=1$	**0.92**	0.93	**0.93**	1.09	**0.92**	1.08	**0.90**	1.06

different rolling window sizes. The forecast performance of the ML estimator, in contrast, varies more with the rolling window size.

Simulation results on forecast accuracy in the high-dimensional designs (for $a = -0.4$) are in Table 4.4. The forecast accuracy of the PML estimator is significantly better than the forecast accuracy of the ML estimator, for all forecast horizons (all p-values< 0.01). The improvements in forecast accuracy are, overall, larger for these high-dimensional designs than for the low-dimensional designs in Table 4.3. The largest gain in forecast accuracy of the PML estimator relative to the ML estimator is attained for the longer forecast horizons.

Table 4.5: *Low-dimensional designs. Frequency of the estimated cointegration rank $\hat{r} = 0, \ldots, q$ using Johansen's trace statistic, the Bartlett-corrected trace statistic, the bootstrap of Cavaliere et al. [2012] and the Rank Selection Criterion (RSC).*

Method \ $\hat{r}$	0	1	2	3	4	0	1	2	3	4
	Sparse $q = 4, T = 500, r = 1$					Sparse $q = 4, T = 500, r = 2$				
Johansen	0.0	95.4	4.2	0.4	0.0	0.0	0.0	96.2	3.8	0.0
Bartlett	0.0	96.0	3.6	0.4	0.0	0.0	0.0	95.4	4.4	0.2
Bootstrap	0.0	96.8	2.8	0.4	0.0	0.0	0.0	97.2	2.8	0.0
RSC	0.0	89.4	10.6	0.0	0.0	0.0	0.0	98.8	1.2	0.0

4.5.4 Rank determination

We evaluate the performance of the Rank Selection Criterion (RSC) in correctly selecting the true cointegration rank. We compare with the trace statistic of Johansen [1988], the Bartlett-corrected trace statistic [Johansen, 2002] and the bootstrap procedure of Cavaliere et al. [2012], where the latter two were proposed to improve the small sample performance of Johansen's trace statistic.[4] For each method, we record the relative frequencies, over all simulation runs, of the selected cointegration ranks.

Results. Table 4.5 reports the results on the cointegration rank estimation for the low-dimensional designs (for $a = -0.4$). In the first sparse setting, the Rank Selection Criterion achieves competitive performance with a rank recovery percentage around 89%. Johansen's method is aimed at controlling size, resulting in a rank recovery percentage around 95% when working with a 5% significance level. Similar results are obtained for the non-sparse low-dimensional setting and are, therefore, omitted. In the second sparse setting, RSC correctly selects the cointegration rank in almost all simulation runs.

Table 4.6 reports the results on the cointegration rank estimation for the high-dimensional designs. In all settings, RSC performs much better than its alternatives. In the first setting, RSC estimates the cointegration rank correctly in 57.4% of the simulation runs, the Bartlett-corrected trace statistic in 11.2%, the bootstrap in 1.2% and Johansen's trace statistic in 0%. Due to the severe size distortions in this small sample size design, the rank recovery percentage of Johansen's trace statistic does not improve when working with a significance level

[4] All tests are conducted at the 5% significance level.

Table 4.6: *High-dimensional designs. Frequency of the estimated cointegration rank* $\hat{r} = 0, \ldots, q$.

Method \ $\hat{r}$	0	1	2	3	4	5	6	7	8	9	10	11
					Sparse $q = 11, T = 50, r = 1$							
Johansen	0.0	0.0	0.0	1.0	9.0	15.2	52.0	14.0	7.0	1.6	0.2	0.0
Bartlett	0.0	11.2	31.8	20.2	14.0	6.6	6.2	4.4	3.8	1.6	0.2	0.0
Bootstrap	98.8	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RSC	0.0	57.4	40.4	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
					Sparse $q = 11, T = 50, r = 4$							
Johansen	0.0	0.0	0.0	3.2	24.6	23.4	41.4	6.4	0.8	0.2	0.0	0.0
Bartlett	0.0	7.6	18.4	23.6	19.0	11.8	10.0	5.4	3.2	0.8	0.2	0.0
Bootstrap	99.0	0.8	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RSC	0.0	0.0	9.0	60.6	28.8	1.6	0.0	0.0	0.0	0.0	0.0	0.0

of, for instance, 1%. Similar results are obtained for the non-sparse setting.

When the true cointegration rank becomes higher ($r = 4$ in the second setting), the performance of the Rank Selection Criterion becomes sensitive to the strength of the cointegration signal: its rank recovery percentage increases from 28.8% for $a = -0.4$ to 73.8% for $a = -0.8$ (unreported). However, even then, RSC is still the best performing method.

In contrast to Johansen's trace statistic, the RSC in not meant to control size. One should take the difficulty in comparing size-targeting methods, such as Johansen's trace statistic, to consistency-targeting methods, such as the RSC, into account when assessing the results on cointegration rank determination. The RSC also has the tendency to overestimate rather than underestimate the cointegration rank. Overestimation is less severe since the PML estimator allows some of the cointegrating vectors, i.e. columns of $\boldsymbol{\beta}$, to be estimated as zero. Then the actual rank of $\hat{\boldsymbol{\beta}}$ will be lower than the estimated rank by the RSC.

4.6 Forecasting

We evaluate the forecast performance of the sparse cointegration method on two data sets. In a first data set, we have interest rates of different maturity. Financial theory implies these interest rates of different maturity to be cointegrated. We consider a VECM and compare the forecast performance of the sparse cointegration method to the traditional method. For the second data set, we forecast a large

number of industry-specific consumption time series. We investigate if forecast accuracy can be improved by using the sparse cointegration method compared to alternative methods.

To evaluate forecast accuracy, we perform rolling window forecasting as described in Section 4.5.3. We use the Rank Selection Criterion from Section 4.4 to estimate the cointegration rank. The BIC criterion is used to select the order p of the VECM. Apart from the Multivariate Mean Absolute Forecast Error, we also provide results for the individual time series $\Delta y_t^{(i)}, i = 1, \ldots, q$ to predict by computing the Mean Absolute Forecast Error

$$\text{MAFE} = \frac{1}{T-h-S+1} \sum_{t=S}^{T-h} \frac{|\Delta y_{t+h}^{(i)} - \widehat{\Delta y}_{t+h}^{(i)}|}{\widehat{\sigma}_{(i)}}. \tag{4.11}$$

To compare forecast performance among different methods, we use the Diebold-Mariano test (DM-test, Diebold and Mariano, 1995).

4.6.1 Interest Rate Growth Forecasting

In finance, the expectations hypothesis of interest rates (e.g. Engsted and Tanggaard, 1994, Giese, 2008) implies interest rates of different maturity to be cointegrated. We collect monthly data on $q = 5$ US treasury bills with different time to maturity (1, 3, 5, 7 and 10 years), ranging from July 1969 until June 2015, hence $T = 552$ (source: Datastream - Federal Reserve, US). A time plot of the interest rates is in Figure 4.1. All interest rates move very closely together, hence, they are expected to be cointegrated. A stationarity test of all individual interest rates using the Augmented Dickey-Fuller test confirms that the time series are integrated of order 1. We take the cointegration relations implied by financial theory into account by estimating a VECM with q interest rates.

We investigate how the Penalized Maximum Likelihood estimator behaves compared to the Johansen Maximum Likelihood estimator when the length of the time series varies relative to the fixed dimension $q = 5$. For this purpose, we consider different window sizes: $S \in \{48, 96, 144\}$.

The Multivariate Mean Absolute Forecast Error is computed for four different forecast horizons (columns) and three different rolling window sizes (rows), see Table 4.7. In all settings, the PML estimator beats Johansen's estimator. A DM-test confirms that, overall, this improvement in forecast performance is significant. The MMAFE of the PML estimator remains relatively stable when the window size varies. The MMAFE of the ML estimator, in contrast, becomes

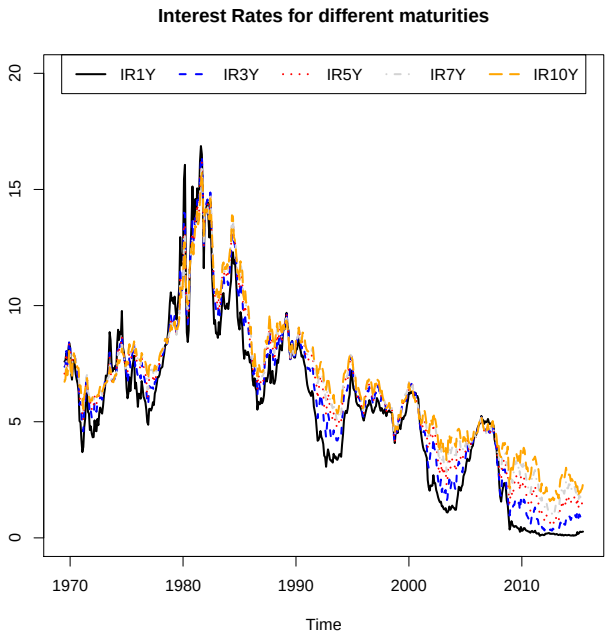


Figure 4.1: *Time plot (July 1969 - June 2015) of the interest rates for the different maturities: 1-year (black solid line), 3-year (blue short dashed line), 5-year (red dotted line), 7-year (gray dotted dashed line), 10-year (orange long dashed line).*

much worse when the window size becomes small. For the window size $S = 48$, the MMAFE of the PML estimator is, on average, 25% lower than the MMAFE of Johansen's estimator. When the window size increases, the PML estimator is still best performing, though the difference between both becomes somewhat smaller.

The Mean Absolute Forecast Errors for the five individual interest rate time series are reported in Table 4.8. The PML estimator delivers the most accurate forecasts, for all interest rates, forecast horizons and window sizes considered. The largest gains in forecast accuracy occur when the window size S is the smallest.

In sum, when the time series length is short compared to the number of time series to predict, important gains in forecast accuracy can be obtained by using the sparse estimator instead of the Johansen's estimator. But even for large time series length, sparsity leads to an improved forecast accuracy for real data, since

Table 4.7: *Multivariate Mean Absolute Forecast Error using the PML and ML estimator. For each window size S (rows) - forecast horizon h (columns) combination, the lowest values are indicated in bold.*

	$h=1$		$h=3$		$h=6$		$h=12$	
Window size	PML	ML	PML	ML	PML	ML	PML	ML
$S=48$	**0.70**	1.13***	**0.70**	0.86***	**0.74**	0.98***	**0.70**	0.86***
$S=96$	**0.68**	0.84**	**0.68**	0.74***	**0.69**	0.77***	**0.70**	0.75**
$S=144$	**0.63**	0.71**	**0.61**	0.66**	**0.59**	0.65***	**0.58**	0.65**

Note: Significance at the 1% (***), 5% (**), and 10% level (*) for the DM-test of equal MMAFE of the two methods.

Table 4.8: *Mean Absolute Forecast Error for the $q=5$ individual interest rate time series using the PML and ML estimator. For each interest rate and window size - forecast horizon combination, the lowest values are indicated in bold.*

Window	Interest	$h=1$		$h=3$		$h=6$		$h=12$	
size	Rate	PML	ML	PML	ML	PML	ML	PML	ML
	1Y	**0.60**	0.73***	**0.61**	0.74***	**0.62**	0.76***	**0.60**	0.66**
	3Y	**0.66**	0.99***	**0.67**	0.86***	**0.68**	0.92***	**0.67**	0.87***
$S=48$	5Y	**0.70**	1.25***	**0.73**	0.92***	**0.77**	1.01***	**0.73**	0.88***
	7Y	**0.73**	1.46***	**0.72**	0.89***	**0.75**	1.04***	**0.74**	0.92***
	10Y	**0.81**	1.19**	**0.79**	0.89*	**0.88**	1.15*	**0.78**	0.97***
	1Y	**0.54**	0.57	**0.55**	0.59**	**0.56**	0.59	**0.55**	0.57
	3Y	**0.66**	0.80**	**0.66**	0.72***	**0.66**	0.72***	**0.68**	0.73**
$S=96$	5Y	**0.70**	0.92**	**0.70**	0.78**	**0.72**	0.83**	**0.70**	0.79***
	7Y	**0.73**	1.01**	**0.73**	0.78***	**0.74**	0.84***	**0.74**	0.79***
	10Y	**0.78**	0.91*	**0.75**	0.84***	**0.78**	0.88**	**0.80**	0.88
	1Y	**0.44**	0.48***	**0.43**	0.45	**0.41**	0.45**	**0.40**	0.43***
	3Y	**0.59**	0.68***	**0.59**	0.65**	**0.57**	0.62**	**0.56**	0.60***
$S=144$	5Y	**0.64**	0.79*	**0.64**	0.71*	**0.63**	0.69***	**0.60**	0.66**
	7Y	**0.69**	0.82*	**0.68**	0.73*	**0.66**	0.73***	**0.64**	0.76*
	10Y	**0.78**	0.78	**0.72**	0.74	**0.69**	0.77**	**0.69**	0.80

Note: Significance at the 1% (***), 5% (**), and 10% level (*) for the DM-test of equal MAFE of the two methods.

the sparse estimator delivers a more parsimonious model.

4.6.2 Consumption Growth Forecasting

Our objective is to predict a large number of industry-specific consumption time series. We collect monthly data on $q = 31$ US consumption time series, ranging from January 1999-April 2015, hence $T = 196$ (see Appendix 4.9 for a data description). Personal consumption accounts for around 70% of GDP in the US and is closely monitored by public policy makers and marketing managers [Fornell et al., 2010]. In contrast to total consumption, industry-specific consumption time series have often been discarded in previous forecasting literature since they are typically highly collinear, which might create estimation problems [Carriero et al., 2011]. We exploit the co-movement between these consumption time series by forecasting total and industry-specific consumption growth in a cointegration framework using the PML estimator from Section 4.3. Time plots of all log-transformed consumption time series are in Figure 4.2, Appendix 4.9. A stationarity test of all individual log-transformed time series using the Augmented Dickey-Fuller test confirms that these time series are integrated of order 1, and we forecast consumption growth in a VECM framework.

We conduct a rolling window forecast exercise using a window of 12 years of data ($S = 144$). We compare the performance of 8 estimators. The first three estimators are estimators for the (log-transformed) consumption time series that account for cointegration. The other estimators are estimators for the consumption growth time series that do not account for cointegration. The estimators are (1) PML estimation of the VECM (cfr. Section 4.3), (2) ML estimation of the VECM, (3) Factor Model of Barogozzi et al. [2016] for non-stationary time series, (4) PML estimation of the VAR, (5) ML estimation of the VAR, (6) the Factor Model of Stock and Watson [2002] for stationary time series, (7) Bayesian estimation of the VAR with the Normal-Inverse Wishart prior introduced in Banbura et al. [2010], and (8) Bayesian Reduced Rank Regression [Carriero et al., 2011], which combines the benefit of rank reduction and Bayesian shrinkage.[5] Note that the forecast performance is always evaluated in terms of MMAFE or MAFE for the time series in *differences*. As a result, forecast errors of the different estimators are comparable. We have included an intercept in the VECM of equation (1) since some of the consumption time series exhibit a drift.

[5] For estimators (3), (7) and (8), the rank and the number of factors k are determined by calculating the maximum eigenvalue ratio criterion $\hat{k}_j = \hat{\lambda}_j / \hat{\lambda}_{j+1}$ for $j = 1, \ldots, q-1$ from the eigenvalues $\hat{\lambda}_j, \ldots, \hat{\lambda}_q$ and selecting $k = \text{argmax}_j \hat{k}_j$.

Table 4.9: *Multivariate Mean Absolute Forecast Error (MMAFE) for the different methods (columns) and forecast horizons h (rows).*

Forecast horizon	Cointegration			No Cointegration				
	PML	ML	Factor Model	PML	ML	Factor Model	Bayesian	Bayesian Reduced Rank
$h = 1$	0.79	0.74	**0.66**	0.94	5.40***	0.72	0.69	0.69
$h = 3$	**0.62**	0.78***	0.66***	0.67***	4.81***	0.75***	0.71***	0.71***
$h = 6$	**0.63**	0.82***	0.67***	0.67***	4.84***	0.77***	0.74***	0.74***
$h = 12$	**0.61**	0.72***	0.65***	0.66***	5.22***	0.72***	0.72***	0.72***

Significance at the 1% (***), 5% (**), and 10% (*) level for the DM-test of equal MMAFE of a given method and the PML method for cointegration.

The Multivariate Mean Forecast Errors are reported in Table 4.9. The PML estimator of the VECM attains the lowest value for all forecast horizons except for $h = 1$.[6] A DM-test confirms that the differences in forecast performance are significant. Taking the long-run cointegration relations into account pays off especially for the longer forecast horizons. Taking cointegration into account (PML, ML, Factor Model) yields in almost all cases significantly lower forecasts compared to not accounting for cointegration. Among the methods that account for cointegration, the PML estimator performs best, confirming that sparse estimation improves forecast performance. The PML estimator of the VECM also performs significantly better than the Bayesian estimators.

Individual Mean Absolute Forecast Errors for the separate time series are also computed. For reasons of brevity, we only report them for the Total consumption time series in Table 4.10. The results for the MAFE are similar to those of the MMAFE. The PML, ML estimator and Factor Model that account for cointegration attain a (significantly) better MAFE than the corresponding methods that do not account for it. The proposed PML estimator of the VECM attains the best value of the MAFE for all forecast horizons except $h = 1$.

In sum, for high-dimensional time series, the sparse cointegration method is a valuable addition to the forecaster's toolbox. It exploits the co-movement between a large number of time series by sparsely estimating the cointegration relations.

[6] Although the Factor Model for cointegration attains the best MMAFE for $h = 1$, its forecast performance is not significantly different from the forecast performance of the PML method for cointegration.

Table 4.10: *Mean Absolute Forecast Error (MAFE) for the Total consumption time series, for different methods (columns) and forecast horizons h (rows).*

Forecast	Cointegration			No Cointegration				
horizon	PML	ML	Factor Model	PML	ML	Factor Model	Bayesian	Bayesian Reduced Rank
$h = 1$	3.82	0.61	**0.59**	6.14	47.28***	0.59	0.65	0.66
$h = 3$	**0.46**	0.66***	0.59***	0.59***	44.44***	0.58*	0.57**	0.57**
$h = 6$	**0.48**	0.81***	0.60**	0.60**	43.96***	0.76***	0.71***	0.71***
$h = 12$	**0.46**	0.62***	0.61***	0.61***	57.61***	0.64**	0.80***	0.79***

See the notes to Table 4.9.

4.7 Conclusion

In this paper, we discuss a sparse cointegration method. Our simulation study shows that the sparse method significantly outperforms Johansen's ML method, if the true cointegrating vectors are sparse or if the time series length is short compared to the number of time series. The degree of sparsity that is needed such that the sparse estimator outperforms the ML estimator depends on the time series length relative to the number of time series. The higher the degree of sparsity, the faster the sparse estimator will outperform the ML estimator.

A sparse cointegration method is useful for several reasons. In high-dimensional settings with cointegrated time series, estimating the cointegrating vectors sparsely might improve estimation accuracy and/or forecast performance. We show that the sparse cointegration method achieves important gains in forecast accuracy compared to the traditional Maximum Likelihood estimator if the time series length is short compared to the number of time series (cfr. interest rate forecasting). When forecasting highly collinear time series (cfr. consumption forecasting), important gains can be obtained by accounting for cointegration and by estimating the cointegration relations sparsely.

The sparse cointegration method might suffer from the following points. We impose the normalization condition on $\boldsymbol{\alpha}$ rather than on $\boldsymbol{\beta}$. As such, the weighted Procrustes problem might be affected by multicollinearity issues. Besides, we impose sparsity on $\boldsymbol{\beta}$, which is not uniquely defined. This might pose difficulties for model interpretation. Their consequences for the forecast performance of the proposed method are, however, less severe.

We use the Rank Selection Criterion of Bunea et al. [2011] to determine the cointegration rank. In high-dimensional simulation settings, the Rank Selection Criterion outperforms Johansen's trace statistic, the Bartlett-corrected

trace statistic and the bootstrap procedure of Cavaliere et al. [2012]. While Johansen's trace statistic is not computable as soon as the total number of lagged time series $(p-1)\cdot q$ exceeds the time series length T, the Rank Selection Criterion as presented in Section 4.4 requires the number of time series q to be smaller than the time series length T. Future research is needed on how to improve its implementation for truly high-dimensional settings where $q > T$. The eigenvalue-ratio-based rank estimator of Lam and Yao [2012] might be an alternative to the RSC for such settings.

There are several questions we did not address, which are left for future research. For instance, the models analyzed in this paper generally exclude deterministic terms [Nielsen and Rahbek, 2000]. We also made abstraction of structural breaks. Allowing for structural breaks is useful when analyzing economic data [Johansen et al., 2000]. A natural extension of this study would be to implement structural analysis. Impulse-response functions, for instance, can be estimated using the PML estimator. Confidence bound around the impulse-response functions are then obtained using a bootstrap procedure.

4.8 Appendix A: Time-series cross-validation

We select the tuning parameters according to a time series cross-validation approach [Hyndman, 2014]. Denote the response by $\mathbf{z}_t$. When solving for $\mathbf{\Gamma}$, $\mathbf{z}_t = \mathbf{\Delta y}_t - \mathbf{\Pi y}_{t-1}$. When solving for $\mathbf{\Pi}$, $\mathbf{z}_t = \mathbf{\Delta y}_t - \sum_{i=1}^{p-1} \mathbf{\Gamma}_i \mathbf{\Delta y}_{t-i}$.

(i) For $t = S, \ldots, T-1$ (with $S = \lfloor 0.8T \rfloor$), repeat:

  (a) For a grid of tuning parameters, fit the model to the data $\mathbf{z}_1, \ldots, \mathbf{z}_t$.

  (b) Compute the one-step-ahead forecast error $\widehat{\mathbf{e}}_{t+1} = \mathbf{z}_{t+1} - \hat{\mathbf{z}}_{t+1}$

(ii) Select the value of the tuning parameter that minimizes the mean squared forecast error

$$\text{MSFE} = \frac{1}{T-S} \sum_{t=S}^{T-1} \frac{1}{q} \sum_{i=1}^{q} \left(\frac{\hat{e}_{t+1}^{(i)}}{\hat{\sigma}_{(i)}} \right)^2 ,$$

with $\hat{e}_t^{(i)}$ the i^{th} component of the multivariate time series at time t and $\hat{\sigma}_{(i)}$ the standard deviation of the time series $z_t^{(i)}$.

4.9 Appendix B: Data description consumption time series

Table 4.11: *Consumption expenditures in billions of US dollars (source: Datastream - Bureau of Economic Analysis).*

Total Consumption
Durable consumption: Motor vehicles and parts
Durable consumption: Furnishings and durable household equipment
Durable consumption: Household appliances
Durable consumption: Recreational goods and vehicles
Durable consumption: Video and Audio equipment
Durable consumption: Photographic equipment
Durable consumption: Information Processing equipment
Durable consumption: Sporting equipment, supplies, guns and ammunition
Durable consumption: Sports and recreational vehicles
Durable consumption: Recreational books
Durable consumption: Musical instruments
Durable consumption: Jewelry
Durable consumption: Watches
Durable consumption: Therapeutic medical equipment
Durable consumption: Corrective eyeglasses and contact lenses
Durable consumption: Educational books
Durable consumption: Luggage
Durable consumption: Telephone equipment
Nondurable Consumption: Food and Beverages
Nondurable Consumption: Food produced and consumed on farms
Nondurable Consumption: Clothing and Footwear
Nondurable Consumption: Gasoline and other energy goods
Nondurable Consumption: Pharmaceutical and Other medical products
Nondurable Consumption: Recreational Items
Nondurable Consumption: Games, Toys and Hobbies
Nondurable Consumption: Flowers, seeds and potted plants
Nondurable Consumption: Film and photographic supplies
Nondurable Consumption: Personal care products
Nondurable Consumption: Magazines and Newspapers
Nondurable Consumption: Net expenditures abroad by US residents

Figure 4.2: *Time plot (January 1999 - April 2015) of the total consumption time series, the 18 durable consumption time series, and the 12 nondurable consumption time series all in logs.*

Chapter 5

Sparse canonical correlation analysis from a predictive point of view

Abstract

Canonical correlation analysis (CCA) describes the associations between two sets of variables by maximizing the correlation between linear combinations of the variables in each data set. However, in high-dimensional settings where the number of variables exceeds the sample size or when the variables are highly correlated, traditional CCA is no longer appropriate. This paper proposes a method for sparse CCA. Sparse estimation produces linear combinations of only a subset of variables from each data set, thereby increasing the interpretability of the canonical variates. We consider the CCA problem from a predictive point of view and recast it into a regression framework. By combining an alternating regression approach together with a lasso penalty, we induce sparsity in the canonical vectors. We compare the performance with other sparse CCA techniques in different simulation settings and illustrate its usefulness on a genomic data set.

5.1 Introduction

The aim of canonical correlation analysis (CCA), introduced by Hotelling [1936], is to identify and quantify linear relations between two sets of variables. CCA is used in various research fields to study associations in, for example, biomedical

data [Foucart, 1999, Alonso et al., 2003], environmental data [Iaci et al., 2010] or genomic data [Graffelman and van Eeuwijk, 2005]. One searches for the linear combinations of each of the two sets of variables having maximal correlation. These linear combinations are called the *canonical variates* and the correlations between the canonical variates are called the *canonical correlations*. We refer to e.g. Johnson and Wichern (1998, Chapter 10) for more information on canonical correlation analysis.

At the same time, we want to induce sparsity in the canonical vectors such that the linear combinations only include a *subset* of the variables. Sparsity is especially helpful in analyzing associations between high-dimensional data sets, which are commonplace today in, for example, genetics [Schwender et al., 2008] and machine learning [Sun et al., 2011, Shin and Wu, 2014]. Therefore, we propose a sparse version of CCA where some elements of the canonical vectors are estimated as exactly zero, which facilitates interpretation. For this aim, we use the formulation of CCA as a prediction problem.

Consider two random vectors $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$. We assume, without loss of generality, that all variables are mean centered and that $p \leq q$. Denote the joint covariance matrix of $(\mathbf{x},\mathbf{y})$ by

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}$$

with $r = \text{rank}(\boldsymbol{\Sigma}_{xy}) \leq p$. Let $\mathbf{A} \in \mathbb{R}^{p \times r}$ and $\mathbf{B} \in \mathbb{R}^{q \times r}$ be the matrices with in their columns the *canonical vectors*. The new variables $\mathbf{u} = \mathbf{A}^\mathbf{T}\mathbf{x}$ and $\mathbf{v} = \mathbf{B}^\mathbf{T}\mathbf{y}$ are the *canonical variates* and the correlations between each pair of canonical variates give the *canonical correlations*. The canonical vectors contained in the matrices $\mathbf{A}$ and $\mathbf{B}$ are respectively given by the eigenvectors of the matrices

$$\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} \quad \text{and} \quad \boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}. \tag{5.1}$$

Both matrices have the same positive eigenvalues, the canonical correlations are given by the positive square root of those eigenvalues.

The canonical vectors and correlations are typically estimated by taking the sample versions of the covariances in (5.1) and computing the corresponding eigenvectors and eigenvalues. However, to implement this procedure, we need to invert the matrices $\widehat{\boldsymbol{\Sigma}}_{xx}$ and $\widehat{\boldsymbol{\Sigma}}_{yy}$. When the original variables are highly correlated or when the number of variables becomes large compared to the sample size, the estimation imprecision will be large. Moreover, when the largest number of variables in both data sets exceeds the sample size n (i.e. $q \geq n$), traditional CCA cannot

be performed since the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{yy}$ is singular, i.e. its inverse does not exist. Vinod [1976] proposed the canonical ridge, which is an adaptation of the ridge regression concept of Hoerl and Kennard [1970] to the framework of CCA, to solve this problem. The canonical ridge replaces the matrices $\widehat{\mathbf{\Sigma}}_{xx}^{-1}$ and $\widehat{\mathbf{\Sigma}}_{yy}^{-1}$ by respectively $(\widehat{\mathbf{\Sigma}}_{xx} + k_1\mathbf{I})^{-1}$ and $(\widehat{\mathbf{\Sigma}}_{yy} + k_2\mathbf{I})^{-1}$. By adding the penalty terms k_1 and k_2 to the diagonal elements of the sample covariance matrices, one obtains more reliable and stable estimates when the data are nearly or exactly collinear.

Another approach is to use sparse CCA techniques. Parkhomenko et al. [2009] consider a sparse singular value decomposition to derive sparse singular vectors. A limitation of their approach is that sparsity in the canonical vectors is only guaranteed if the variables within the first data set and the variables within the second data set are uncorrelated. A similar approach was taken by Witten and Tibshirani [2009] who apply a penalized matrix decomposition to the cross-product matrix $\widehat{\mathbf{\Sigma}}_{xy}$, but they also require uncorrelatedness of the variables within each of the two data sets. Waaijenborg et al. [2008] consider Wold's (1968) alternating least squares approach to CCA and obtain sparse canonical vectors using penalized regression with the elastic net. The ridge parameter of the elastic net is set to be large, thereby, according to the authors, ignoring the dependency structure within each set of variables.

Waaijenborg et al. [2008], Witten and Tibshirani [2009], and Parkhomenko et al. [2009] all require the variables within each of the two data sets to be uncorrelated. This uncorrelatedness restriction is restrictive since data sets containing correlated variables are commonplace in multivariate analysis (e.g. genome-wide association studies). Therefore, we propose in this paper to estimate the canonical variates without imposing any prior covariance restrictions. As soon as the data sets contain correlated variables, the gains in estimation accuracy achieved by our sparse CCA method compared to these three other sparse CCA methods are outspoken.

We consider CCA as a prediction problem, where the canonical variates obtained from the first data set serve as optimal predictors for the canonical variates of the second data set, and vice versa. Our proposed method obtains the canonical vectors using an alternating penalized regression framework. By performing variable selection in a penalized regression framework using the lasso penalty [Tibshirani, 1996], we obtain sparse canonical vectors. We demonstrate in a simulation study that our Sparse Alternating Regression (SAR) algorithm produces good results in terms of estimation accuracy of the canonical vectors, and detec-

tion of the sparseness structure of the canonical vectors. We also apply the SAR algorithm to a high-dimensional genomic data set. Sparse estimation is appealing since it highlights the most important variables for the association study.

The remainder of this article is organized as follows. In Section 5.2 we formulate the CCA problem from a predictive point of view. Section 5.3 describes the Sparse Alternating Regression (SAR) approach and provides details on the implementation of the algorithm. Section 5.4 compares our methodology to other sparse CCA techniques by means of a simulation study. Section 5.5 discusses the genomic data example, Section 5.6 concludes.

5.2 CCA from a predictive point of view

A characterization of the canonical vectors based on the concept of prediction is proposed by Brillinger [1975] and Izenman [1975]. Given n observations $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{y}_i \in \mathbb{R}^q$ $(i = 1, \ldots, n)$, consider the optimization problem

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \underset{(\mathbf{A},\mathbf{B})\in\mathcal{S}}{\operatorname{argmin}} \sum_{i=1}^{n} ||\mathbf{A}^T\mathbf{x}_i - \mathbf{B}^T\mathbf{y}_i||^2. \tag{5.2}$$

We restrict the parameter space to the space $\mathcal{S}$, given by

$$\mathcal{S} = \{(\mathbf{A}, \mathbf{B}) : \mathbf{A} \in \mathbb{R}^{p\times r}, \mathbf{B} \in \mathbb{R}^{q\times r}, \operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{B}) = r, \mathbf{A^T\Sigma}_{xx}\mathbf{A} = \mathbf{B^T\Sigma}_{yy}\mathbf{B} = \mathbf{I}_r\}.$$

We impose normalization conditions requiring the canonical variates to have unit variance and to be uncorrelated. Brillinger [1975] proves that the objective function in (6.1) is minimized when $\mathbf{A}$ and $\mathbf{B}$ contain in their columns the canonical vectors.

We build on this equivalent formulation of the CCA problem to obtain the canonical vectors using an alternating regression procedure (see e.g. Wold, 1968; Branco et al., 2005). The subsequent canonical variates are sequentially derived. Since we consider CCA in a regression framework, we do not have to estimate the covariance matrices of equation (5.1). Furthermore, normality of the data is not required.

First canonical vector pair. Denote the first canonical vectors (i.e. the first columns of the matrices $\mathbf{A}$ and $\mathbf{B}$) by $(\mathbf{A}_1, \mathbf{B}_1)$. Suppose we have an initial value $\mathbf{A}_1$ for the first canonical vector in the matrix $\mathbf{A}$. Then the minimization problem in (6.1) reduces to

$$\widehat{\mathbf{B}}_1|\mathbf{A}_1 = \underset{\mathbf{B}_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\mathbf{A}_1^T\mathbf{x}_i - \mathbf{B}_1^T\mathbf{y}_i\right)^2, \tag{5.3}$$

where we require $\hat{\mathbf{v}}_1 = \mathbf{Y}\widehat{\mathbf{B}}_1$ to have unit variance. The solution to (6.2) can be obtained from a multiple regression with $\mathbf{X}\mathbf{A}_1^*$ as response and $\mathbf{Y}$ as predictor, where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$.

We proceed analogously for fixed value $\mathbf{B}_1$. The optimal value for $\mathbf{A}_1$ is obtained by a multiple regression with $\mathbf{Y}\mathbf{B}_1$ as response and $\mathbf{X}$ as predictor

$$\widehat{\mathbf{A}}_1|\mathbf{B}_1 = \underset{\mathbf{A}_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\mathbf{B}_1^T\mathbf{y}_i - \mathbf{A}_1^T\mathbf{x}_i\right)^2, \tag{5.4}$$

where we require $\hat{\mathbf{u}}_1 = \mathbf{X}\widehat{\mathbf{A}}_1$ to have unit variance. This leads to an alternating regression scheme, where we alternately update our estimates of the first canonical vectors until convergence. We iterate until the relative change in the value of the objective function in two successive iterations is smaller than the convergence tolerance value $\epsilon = 10^{-2}$.

Higher order canonical vector pairs. The higher order canonical variates need to be orthogonal to the previously found canonical variates. Therefore, the alternating regression scheme is applied to deflated data matrices (see e.g. Branco et al., 2005). For the second pair of canonical vectors, consider the deflated matrices

$$\mathbf{X}_2^* = \mathbf{X} - \hat{\mathbf{u}}_1(\hat{\mathbf{u}}_1^T\hat{\mathbf{u}}_1)^{-1}\hat{\mathbf{u}}_1^T\mathbf{X}. \tag{5.5}$$

The deflated matrix $\mathbf{X}_2^*$ is obtained as the residuals of the multivariate regression of $\mathbf{X}$ on $\hat{\mathbf{u}}_1$, the first canonical variate. Analogously, the deflated matrix $\mathbf{Y}_2^*$ is given by

$$\mathbf{Y}_2^* = \mathbf{Y} - \hat{\mathbf{v}}_1(\hat{\mathbf{v}}_1^T\hat{\mathbf{v}}_1)^{-1}\hat{\mathbf{v}}_1^T\mathbf{Y}, \tag{5.6}$$

the residuals of the multivariate regression of $\mathbf{Y}$ on $\hat{\mathbf{v}}_1$.

Using the Least Squares property, each column of $\mathbf{X}_2^*$ is uncorrelated with the first canonical variate $\hat{\mathbf{u}}_1$. The second canonical variate will be a linear combination of the columns of $\mathbf{X}_2^*$ and, hence, will be uncorrelated to the previously found canonical variate. The same holds for $\mathbf{Y}_2^*$. The second canonical variate pair is then obtained by alternating between the following regressions until convergence:

$$\widehat{\mathbf{B}}_2^*|\mathbf{A}_2^* = \underset{\mathbf{B}_2^*}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\mathbf{A}_2^{*T}\mathbf{x}_{2,i}^* - \mathbf{B}_2^{*T}\mathbf{y}_{2,i}^*\right)^2, \tag{5.7}$$

$$\widehat{\mathbf{A}}_2^*|\mathbf{B}_2^* = \underset{\mathbf{A}_2^*}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\mathbf{B}_2^{*T}\mathbf{y}_{2,i}^* - \mathbf{A}_2^{*T}\mathbf{x}_{2,i}^*\right)^2, \tag{5.8}$$

where we require $\hat{\mathbf{v}}_2^* = \mathbf{Y}_2^*\widehat{\mathbf{B}}_2^*$ and $\hat{\mathbf{u}}_2^* = \mathbf{X}_2^*\widehat{\mathbf{A}}_2^*$ to have both unit variance.

Finally, we need to express the second canonical vector pair in terms of the original data sets $\mathbf{X}$ and $\mathbf{Y}$. To obtain the second canonical vector $\widehat{\mathbf{A}}_2$, we regress $\hat{\mathbf{u}}_2^*$ on $\mathbf{X}$

$$\widehat{\mathbf{A}}_2 = \underset{\mathbf{A}_2}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\hat{\mathbf{u}}_{2,i}^* - \mathbf{A}_2^T \mathbf{x}_i\right)^2, \tag{5.9}$$

yielding the fitted values $\hat{\mathbf{u}}_2 = \mathbf{X}\widehat{\mathbf{A}}_2$. To obtain $\widehat{\mathbf{B}}_2$, we regress $\hat{\mathbf{v}}_2^*$ on $\mathbf{Y}$.

$$\widehat{\mathbf{B}}_2 = \underset{\mathbf{B}_2}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\hat{\mathbf{v}}_{2,i}^* - \mathbf{B}_2^T \mathbf{y}_i\right)^2. \tag{5.10}$$

The same idea is applied to obtain the higher order canonical variate pairs.

5.3 Sparse alternating regressions

The canonical vectors obtained with the alternating regression scheme from Section 5.2 are in general not sparse. Sparse canonical vectors are obtained by replacing the Least Squares regressions in the alternating regression approach of Section 5.2 with Lasso regressions (L_1-penalty). In contrast to Ridge regressions (L_2-penalty), the constraint region of the Lasso is such that an estimated regression coefficients will sometimes be set to exactly zero (Tibshirani, 1996; Figure 2). As such, some coefficients in the canonical vectors will be set to exactly zero, thereby producing linear combinations of only a subset of variables.

For the first pair of sparse canonical vectors, the sparse equivalents of the Least Squares regressions in equations (6.2) and (6.3) are given by

$$\widehat{\mathbf{B}}_1 | \mathbf{A}_1 = \underset{\mathbf{B}_1}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\mathbf{A}_1^T \mathbf{x}_i - \mathbf{B}_1^T \mathbf{y}_i\right)^2 + \lambda_{\mathbf{B}_1} \sum_{j=1}^{q} |b_{j1}| \right),$$

$$\widehat{\mathbf{A}}_1 | \mathbf{B}_1 = \underset{\mathbf{A}_1}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\mathbf{B}_1^T \mathbf{y}_i - \mathbf{A}_1^T \mathbf{x}_i\right)^2 + \lambda_{\mathbf{A}_1} \sum_{j=1}^{p} |a_{j1}| \right),$$

where $\lambda_{\mathbf{B}_1} > 0$ and $\lambda_{\mathbf{A}_1} > 0$ are sparsity parameters, b_{j1} is the j^{th} $(j = 1, \ldots, q)$ element of the first canonical vector $\mathbf{B}_1$ and a_{j1} is the j^{th} $(j = 1, \ldots, p)$ element of the first canonical vector $\mathbf{A}_1$. The first pair of canonical variates are given by $\hat{\mathbf{u}}_1 = \mathbf{X}\widehat{\mathbf{A}}_1$ and $\hat{\mathbf{v}}_1 = \mathbf{Y}\widehat{\mathbf{B}}_1$. We require both to have unit variance.

To obtain the second pair of sparse canonical vectors, the same deflated matrices as in equations (5.5) and (5.6) are used. The Least Squares regressions in

equations (5.7) and (5.8) are replaced by the Lasso regressions

$$\widehat{\mathbf{B}}_2^*|\mathbf{A}_2^* = \underset{\mathbf{B}_2^*}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\mathbf{A}_2^{*T}\mathbf{x}_{2,i}^* - \mathbf{B}_2^{*T}\mathbf{y}_{2,i}^* \right)^2 + \lambda_{\mathbf{B}_2^*} \sum_{j=1}^{q} |b_{j2}^*| \right),$$

$$\widehat{\mathbf{A}}_2^*|\mathbf{B}_2^* = \underset{\mathbf{A}_2^*}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\mathbf{B}_2^{*T}\mathbf{y}_{2,i}^* - \mathbf{A}_2^{*T}\mathbf{x}_{2,i}^* \right)^2 + \lambda_{\mathbf{A}_2^*} \sum_{j=1}^{p} |a_{j2}^*| \right).$$

Finally, to express the second pair of canonical vectors in terms of the original data matrices, we replace the Least Squares regression in (5.9) and (5.10) by the two Lasso regressions.

$$\widehat{\mathbf{A}}_2 = \underset{\mathbf{A}_2}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\hat{\mathbf{u}}_{2,i}^* - \mathbf{A}_2^T\mathbf{x}_i \right)^2 + \lambda_{\mathbf{A}_2} \sum_{j=1}^{p} |a_{j2}| \right),$$

$$\widehat{\mathbf{B}}_2 = \underset{\mathbf{B}_2}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\hat{\mathbf{v}}_{2,i}^* - \mathbf{B}_2^T\mathbf{y}_i \right)^2 + \lambda_{\mathbf{B}_2} \sum_{j=1}^{q} |b_{j2}| \right),$$

yielding the fitted values $\hat{\mathbf{u}}_2 = \mathbf{X}\widehat{\mathbf{A}}_2$ and $\hat{\mathbf{v}}_2 = \mathbf{Y}\widehat{\mathbf{B}}_2$. We add a lasso penalty to the above regressions, first because the design matrix $\mathbf{X}$ can be high-dimensional, and second, because we want $\widehat{\mathbf{A}}_2$ and $\widehat{\mathbf{B}}_2$ to be sparse.

A complete description of the Sparse Alternating Regression (SAR) algorithm is given below. We numerically verified that without imposing penalization (i.e. $\lambda_{\mathbf{A}_l^*} = \lambda_{\mathbf{B}_l^*} = \lambda_{\mathbf{A}_l} = \lambda_{\mathbf{B}_l} = 0$, for $l = 1, \ldots, r$), the traditional CCA solution is obtained. Finally, note that as in other sparse CCA proposals (Witten and Tibshirani, 2009; Parkhomenko et al., 2009; Waaijenborg et al., 2008) the sparse canonical variates are in general not uncorrelated. We do not consider this lack of uncorrelatedness as a flaw. The sparse canonical vectors yield an easily interpretable basis of the space spanned by the canonical vectors. After suitable rotation of the corresponding canonical variates, this basis can be made orthogonal (but not sparse) if one desires so.

Sparse Alternating Regression (SAR) Algorithm

Let $\mathbf{X}$ and $\mathbf{Y}$ be two data matrices.

(i) *Preliminary steps*

- $\mathbf{X}_0 := \mathbf{X}_1^* = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T$
- $\mathbf{Y}_0 := \mathbf{Y}_1^* = \mathbf{Y} - \mathbf{1}\bar{\mathbf{y}}^T$

(ii) *Alternating Regressions*: For $l = 1, \ldots, r$

- If $l > 1$: *Deflated matrices*
  - $\mathbf{X}_l^* = \mathbf{X} - \widehat{\mathbf{U}}_l(\widehat{\mathbf{U}}_l^T\widehat{\mathbf{U}}_l)^{-1}\widehat{\mathbf{U}}_l^T\mathbf{X}$, with $\widehat{\mathbf{U}}_l = [\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_{l-1}]$
  - $\mathbf{Y}_l^* = \mathbf{Y} - \widehat{\mathbf{V}}_l(\widehat{\mathbf{V}}_l^T\widehat{\mathbf{V}}_l)^{-1}\widehat{\mathbf{V}}_l^T\mathbf{Y}$, with $\widehat{\mathbf{V}}_l = [\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_{l-1}]$
- *Starting values*
  - $\widehat{\mathbf{B}}_l^{(0)} = \frac{\hat{b}_l^{\text{can ridge}}}{||\hat{b}_l^{\text{can ridge}}||}$, using the canonical vector $\hat{b}_l^{\text{can ridge}}$ obtained with the canonical ridge. Regularization parameters are chosen using 5-fold cross-validation such that the average test sample canonical correlation is maximized.
  - $\widehat{\mathbf{v}}_l^{*(0)} = \mathbf{Y}_l^*\widehat{\mathbf{B}}_l^{(0)}$
- *From iteration* $s = 1$ *until convergence.Sparsity parameters selected using BIC (cfr. Section 5.3).*
  - $$\widehat{\mathbf{A}}_l^{*(s)} = \underset{\mathbf{A}_l^*}{\text{argmin}} \left(\sum_{i=1}^{n} \left(\widehat{\mathbf{v}}_{l,i}^{*(s-1)} - \mathbf{x}_{l,i}^{*T}\mathbf{A}_l^*\right)^2 + \lambda_{\mathbf{A}_l^*} \sum_{j=1}^{p} |a_{jl}^*| \right) \tag{5.11}$$
  - $\widehat{\mathbf{A}}_l^{*(s)} = \frac{\widehat{\mathbf{A}}_l^{*(s)}}{||\widehat{\mathbf{A}}_l^{*(s)}||}$
  - $\widehat{\mathbf{u}}_l^{*(s)} = \mathbf{X}_l^*\widehat{\mathbf{A}}_l^{*(s)}$
  - $$\widehat{\mathbf{B}}_l^{*(s)} = \underset{\mathbf{B}_l^*}{\text{argmin}} \left(\sum_{i=1}^{n} \left(\widehat{\mathbf{u}}_{l,i}^{*(s)} - \mathbf{y}_{l,i}^{*T}\mathbf{B}_l^*\right)^2 + \lambda_{\mathbf{B}_l^*} \sum_{j=1}^{q} |b_{jl}^*| \right) \tag{5.12}$$
  - $\widehat{\mathbf{B}}_l^{*(s)} = \frac{\widehat{\mathbf{B}}_l^{*(s)}}{||\widehat{\mathbf{B}}_l^{*(s)}||}$
  - $\widehat{\mathbf{v}}_l^{*(s)} = \mathbf{Y}_l^*\widehat{\mathbf{B}}_l^{*(s)}$
- *After convergence*, resulting in $\widehat{\mathbf{A}}_l^*, \widehat{\mathbf{B}}_l^*, \widehat{\mathbf{u}}_l^*$ and $\widehat{\mathbf{v}}_l^*$
  - $$\widehat{\mathbf{A}}_l = \begin{cases} \widehat{\mathbf{A}}_l^* & \text{if } l = 1 \\ \underset{\mathbf{A}_l}{\text{argmin}} \left(\sum_{i=1}^{n} (\widehat{\mathbf{u}}_{l,i}^* - \mathbf{x}_{0,i}^T\mathbf{A}_l)^2 + \lambda_{\mathbf{A}_l} \sum_{j=1}^{p} |a_{jl}| \right) & \text{if } l > 1 \end{cases} \tag{5.13}$$
  - $\widehat{\mathbf{u}}_l = \mathbf{X}_0\widehat{\mathbf{A}}_l$
  - $$\widehat{\mathbf{B}}_l = \begin{cases} \widehat{\mathbf{B}}_l^* & \text{if } l = 1 \\ \underset{\mathbf{B}_l}{\text{argmin}} \left(\sum_{i=1}^{n} (\widehat{\mathbf{v}}_{l,i}^* - \mathbf{y}_{0,i}^T\mathbf{B}_l)^2 + \lambda_{\mathbf{B}_l} \sum_{j=1}^{q} |b_{jl}| \right) & \text{if } l > 1 \end{cases} \tag{5.14}$$
  - $\widehat{\mathbf{v}}_l = \mathbf{Y}_0\widehat{\mathbf{B}}_l$
- *Final solution*
  - $\widehat{\mathbf{A}}_{\text{sparse}} = [\widehat{\mathbf{A}}_1, \ldots, \widehat{\mathbf{A}}_r]$
  - $\widehat{\mathbf{B}}_{\text{sparse}} = [\widehat{\mathbf{B}}_1, \ldots, \widehat{\mathbf{B}}_r]$

Optimization problem. We obtain the canonical vectors using a sequential algorithm. The sequential algorithm corresponds to the following sequentially defined optimization criteria.

$$(\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1) = \underset{(\mathbf{A}_1, \mathbf{B}_1)}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\mathbf{A}_1^T \mathbf{x}_i - \mathbf{B}_1^T \mathbf{y}_i \right)^2 + \lambda_{\mathbf{A}_1} \sum_{j=1}^{p} |a_{j1}| + \lambda_{\mathbf{B}_1} \sum_{j=1}^{q} |b_{j1}| \right)$$

$$(\widehat{\mathbf{A}}_l^*, \widehat{\mathbf{B}}_l^*) = \underset{(\mathbf{A}_l^*, \mathbf{B}_l^*)}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\mathbf{A}_l^{*T} \mathbf{x}_{l,i}^* - \mathbf{B}_l^{*T} \mathbf{y}_{l,i}^* \right)^2 + \lambda_{\mathbf{A}_l^*} \sum_{j=1}^{p} |a_{jl}^*| + \lambda_{\mathbf{B}_l^*} \sum_{j=1}^{q} |b_{jl}^*| \right),$$

for $l = 2, \ldots, r$, with the deflated data matrices

$$\mathbf{X}_l^* = \mathbf{X} - \widehat{\mathbf{U}}_l (\widehat{\mathbf{U}}_l^T \widehat{\mathbf{U}}_l)^{-1} \widehat{\mathbf{U}}_l^T \mathbf{X}$$

$$\mathbf{Y}_l^* = \mathbf{Y} - \widehat{\mathbf{V}}_l (\widehat{\mathbf{V}}_l^T \widehat{\mathbf{V}}_l)^{-1} \widehat{\mathbf{V}}_l^T \mathbf{Y},$$

where $\widehat{\mathbf{U}}_l = [\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_{l-1}]$ and $\widehat{\mathbf{V}}_l = [\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_{l-1}]$.

Starting values. To start up the SAR algorithm, an initial value is required. We use the canonical vectors delivered by the canonical ridge as starting value, which is available at no computational cost. The regularization parameters of the canonical ridge are chosen using 5-fold cross-validation such that the average test sample canonical correlation is maximized [Gonzalez et al., 2008].

Number of canonical variates to extract. For practical implementation, one needs to have an idea on the number of canonical variates r to extract. Most often, only a limited number of canonical variate pairs are truly relevant. We follow An et al. [2013] who propose the maximum eigenvalue ratio criterion to decide on the number of canonical variates to extract. We apply the canonical ridge and calculate the canonical correlations $\hat{\rho}_1, \ldots, \hat{\rho}_{\text{rmax}}$, with $\text{rmax} = \min(p, q, 10)$. Let $\hat{k}_j = \hat{\rho}_j / \hat{\rho}_{j+1}$ for $j = 1, \ldots, \text{rmax} - 1$. Then we set $r = \operatorname{argmax}_j \hat{k}_j$, and extract r pairs of canonical variates using the SAR algorithm.

Selection of sparsity parameters. In the SAR algorithm, the sparsity parameters $\lambda_{\mathbf{A}_l^*}$ in equation (5.11) and $\lambda_{\mathbf{B}_l^*}$ in equation (5.12), which control the penalization on the respective regression coefficient matrices, need to be selected. We select the sparsity parameters according to a minimal Bayes Information Criterion (BIC). BIC shows good performance in selecting the tuning parameters (see e.g. Yin and Li, 2011). Moreover, BIC requires less computation time than, for instance, cross-validation.

We solve the corresponding penalized regression problems over a range of values and select for each the one with lowest value of

$$\text{BIC}_{\lambda_{\mathbf{A}_l^*}} = n \cdot \log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{B}_l^{*T}\mathbf{y}_{l,i}^* - \mathbf{A}_l^{*T}\mathbf{x}_{l,i}^*\right)^2\right) + df_{\lambda_{\mathbf{A}_l^*}} \cdot \log(n)$$

$$\text{BIC}_{\lambda_{\mathbf{B}_l^*}} = n \cdot \log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{A}_l^{*T}\mathbf{x}_{l,i}^* - \mathbf{B}_l^{*T}\mathbf{y}_{l,i}^*\right)^2\right) + df_{\lambda_{\mathbf{B}_l^*}} \cdot \log(n),$$

for $l = 1, \ldots, r$, and with $df_{\lambda_{\mathbf{A}_l^*}}$ the number of non-zero estimated regression coefficients. We proceed analogously for $\lambda_{\mathbf{B}_l^*}$. We work with two BIC criteria since we sequentially select the sparsity parameters $\lambda_{\mathbf{A}_l^*}$ (when solving for $\mathbf{A}_l^*$ conditional on $\mathbf{B}_l^*$) and $\lambda_{\mathbf{B}_l^*}$ (when solving for $\mathbf{B}_l^*$ conditional on $\mathbf{A}_l^*$), for $l = 1, \ldots, r$. We proceed analogously to select the sparsity parameters $\lambda_{\mathbf{A}_l}$ and $\lambda_{\mathbf{B}_l}$ for the original variables in respectively equations (5.13) and (5.14).

5.4 Simulation Study

We compare the performance of the Sparse Alternating Regression approach with three other sparse CCA techniques. We consider

(i) The Sparse Alternating Regression (SAR) algorithm detailed in Section 5.3.

(ii) The sparse CCA of Witten and Tibshirani (2009; Available in the `R package PMA`, see Witten et al., 2011), relying on a penalized matrix decomposition applied to the cross-product matrix $\widehat{\mathbf{\Sigma}}_{xy}$. Sparsity parameters are selected using the permutation approach described in Gross et al. [2011].

(iii) The sparse CCA of Parkhomenko et al. (2009; Available at `http://www.uhnres.utoronto.ca/labs/tritchler/`). Sparsity parameters are selected using 5-fold cross-validation where the average test sample canonical correlation is maximized.

(iv) The sparse CCA of Waaijenborg et al. [2008]. The lasso parameter of the elastic net is selected using 5-fold cross-validation such that the mean absolute difference between the canonical correlation of the training and test sets is minimized. We re-implemented the algorithm of Waaijenborg et al. [2008] in `R`.

We emphasize that the sparsity parameters of all methods are selected as proposed by the respective authors. As a robustness check and for more fair comparison

between the methods, we also compare the performance when the tuning parameters are selected in a consistent way across the competing approaches. For this purpose, we use a 5-fold cross-validation where the average test sample canonical correlation is maximized. The traditional CCA solution and the canonical ridge (Available in the `R package CCA`, see Gonzalez and Dejean, 2009) are computed as additional benchmarks.

We consider several simulation designs. For each setting we generate data matrices $\mathbf{X}$ and $\mathbf{Y}$ according to multivariate normal distributions, with covariance matrices described in Table 5.1. In all simulation settings except for the Non-Sparse High-dimensional design, the canonical vectors have a sparse structure. In the 'Uncorrelated' and 'Noisy' design (revised from Branco et al., 2005) the uncorrelatedness restriction of Waaijenborg et al. [2008], Witten and Tibshirani [2009] and Parkhomenko et al. [2009] is satisfied. This restriction is violated in the other simulation designs. In the 'Noisy' design, we investigate the influence of adding a noise term δ in the data generating process. The true canonical vectors in the 'Noisy' and 'Uncorrelated' design are the same, the true canonical correlations are weaker in the 'Noisy design. In the 'Sparse High-dimensional' and the 'NonSparse High-dimensional' design, the number of variables is large compared to the sample size. Traditional CCA can still be performed in this setting. In the 'UltraHigh-dimensional' design, the number of variables in the data matrix $\mathbf{Y}$ is much larger than the sample size, and traditional CCA can no longer be performed. The number of simulations for each setting except for the UltraHigh-dimensional design is $M = 1000$. For the UltraHigh-dimensional design $M = 200$.

5.4.1 Performance measures

We compare the SAR algorithm to its alternatives and evaluate (i) the accuracy of the space spanned by the estimated canonical vectors, and (ii) the detection of the sparsity structure of the canonical vectors.

(i) We compute for each simulation run m, with $m = 1, \ldots, M$, the angle $\theta^m(\hat{\mathbf{A}}^{(m)}, \mathbf{A})$ between the subspace spanned by the estimated canonical vectors contained in the columns of $\hat{\mathbf{A}}^{(m)}$ and the subspace spanned by the true canonical vectors contained in the columns of $\mathbf{A}$. We proceed analogously for the matrix $\mathbf{B}$. The average angles, measuring the accuracy, are given by

$$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A}) = \frac{1}{M} \sum_{m=1}^{M} \theta^m(\hat{\mathbf{A}}^{(m)}, \mathbf{A}) \quad \text{and} \quad \bar{\theta}(\hat{\mathbf{B}}, \mathbf{B}) = \frac{1}{M} \sum_{m=1}^{M} \theta^m(\hat{\mathbf{B}}^{(m)}, \mathbf{B}).$$

(ii) We monitor the sparsity recognition performance (e.g. Rothman et al.,

Table 5.1: *Simulation settings.*

Design	$\mathbf{\Sigma_{xx}}$	$\mathbf{\Sigma_{yy}}$	$\mathbf{\Sigma_{xy}}$
Uncorrelated $n = 50, p = 4, q = 6$	$\mathbf{I_p}$	$\mathbf{I_q}$	$\begin{bmatrix} \frac{3}{5} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$
Noisy $n = 50, p = 4, q = 6$	$\mathbf{I_p} + \delta\mathbf{I_p}$ with $\delta = 0.5$	$\mathbf{I_q} + \delta\mathbf{I_q}$ with $\delta = 0.5$	$\begin{bmatrix} \frac{3}{5} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$
Correlated $n = 50, p = 6, q = 10$	$\mathbf{I_p}$	$\begin{bmatrix} \mathbf{S_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I_7} \end{bmatrix}$ with $\mathbf{S_1}_{ij} = 0.7^{\lvert i-j \rvert}$	$\begin{bmatrix} \mathbf{D_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ with $\mathbf{D_1} = \frac{1}{2}\mathbf{I_2}$
Sparse High-dimensional $n = 50, p = 25, q = 40$	$\mathbf{I_p}$	$\begin{bmatrix} \mathbf{S_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I_{37}} \end{bmatrix}$ with $\mathbf{S_1}_{ij} = 0.3^{\lvert i-j \rvert}$	$\begin{bmatrix} \mathbf{D_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ with $\mathbf{D_1} = \frac{7}{10}\mathbf{I_2}$
NonSparse High-dimensional $n = 50, p = 25, q = 40$	$\mathbf{S_1}$ with $\mathbf{S_1}_{ij} = \begin{cases} 1 & i = j \\ 0.1 & i \neq j \end{cases}$	$\mathbf{S_1}$ with $\mathbf{S_1}_{ij} = \begin{cases} 1 & i = j \\ 0.1 & i \neq j \end{cases}$	$\begin{bmatrix} 0.9 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$
UltraHigh-dimensional $n = 50, p = 1000, q = 1000$	$10^{-2} \cdot \begin{bmatrix} \mathbf{S_1} & \mathbf{0_{9\times 991}} \\ \mathbf{0_{991\times 9}} & \mathbf{I_{991}} \end{bmatrix}$ with $\mathbf{S_1} = \begin{bmatrix} \mathbf{S_3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S_3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S_3} \end{bmatrix}$ $\mathbf{S_3}_{ij} = \begin{cases} 1 & i = j \\ 0.9 & (i,j) = \{(1,2),(2,1)\} \\ 0 & \text{otherwise} \end{cases}$	$10^{-2} \cdot \begin{bmatrix} \mathbf{S_2} & \mathbf{0_{9\times 991}} \\ \mathbf{0_{991\times 9}} & \mathbf{I_{991}} \end{bmatrix}$ with $\mathbf{S_2} = \begin{bmatrix} \mathbf{S_3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S_3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S_3} \end{bmatrix}$ $\mathbf{S_3}_{ij} = \begin{cases} 1 & i = j \\ 0.8 & i \neq j \end{cases}$	$10^{-2} \cdot \mathbf{D_1}$ with $\mathbf{D_1}_{ij} = \begin{cases} 0.2 & i = j = \{1,4,7\} \\ 0 & \text{otherwise} \end{cases}$

2010) using the true positive rate and the true negative rate as defined as follows

$$\text{TPR}(\hat{\mathbf{A}}, \mathbf{A}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\#\{(i,j) : \widehat{\mathbf{A}}_{ij}^{(m)} \neq 0 \text{ and } \mathbf{A}_{ij} \neq 0\}}{\#\{(i,j) : \mathbf{A}_{ij} \neq 0\}}$$
$$\text{TNR}(\hat{\mathbf{A}}, \mathbf{A}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\#\{(i,j) : \widehat{\mathbf{A}}_{ij}^{(m)} = 0 \text{ and } \mathbf{A}_{ij} = 0\}}{\#\{(i,j) : \mathbf{A}_{ij} = 0\}}.$$

The true positive rate indicates the number of true relevant variables detected by the estimation procedure. The true negative rate measures the hit rate of excluding unimportant variables from the canonical vectors. Analogue measures can be computed for the canonical vectors in the matrix $\mathbf{B}$.

5.4.2 Results

The simulation results on the estimation accuracy of the estimated canonical vectors are reported in Table 5.2. We compute the average angle (averaged across simulation runs) between the space spanned by the true and estimated canonical vectors. To compare the average angle of the SAR algorithm against the other approaches, we compute p-values of a two-sided paired t-test. We first discuss the performance of the CCA methods when the tuning parameters are selected using the authors' method.

We first compare the performance of the penalized CCA techniques (i.e. canonical ridge and sparse CCA) to the unpenalized CCA solution. The estimation accuracy of the penalized CCA methods is significantly better compared to traditional CCA, especially in the high-dimensional designs. Interestingly, even in the NonSparse High-dimensional design, the penalized CCA methods perform much better than CCA. As the number of variables approaches the sample size, the estimation imprecision of CCA becomes large. Imposing regularization either via the canonical ridge of via sparse CCA improves estimation accuracy considerably. In the lower dimensional simulation settings (i.e. Uncorrelated, Noisy and Correlated design), sparse CCA techniques are still doing well since the underlying structure of the canonical vectors is sparse. For all methods, estimation accuracy is lower in the Noisy design compared to the Uncorrelated design. The true canonical correlations are weaker in the Noisy design, thus, the signal is weaker, making it more difficult to estimate the canonical vectors. The relative ranking in performance of the different methods remains the same when the same selection method for the sparsity parameters is used compared to the authors' method. It should be noted that the canonical ridge and CCA show very good performance

Table 5.2: *Estimation accuracy of the canonical vectors, measured by the average angle between the subspace spanned by the true and estimated canonical vectors. P-values comparing SAR to alternatives are all* < 0.01*, except for the ones reported in parentheses.*

Design	Method	Tuning parameters selected using			
		Authors' method		Cross-validation	
		$\theta(\hat{\mathbf{A}}, \mathbf{A})$	$\theta(\hat{\mathbf{B}}, \mathbf{B})$	$\theta(\hat{\mathbf{A}}, \mathbf{A})$	$\theta(\hat{\mathbf{B}}, \mathbf{B})$
Uncorrelated	SAR	0.011	0.022	0.036	0.066
	Witten and Tibshirani [2009]	0.010 (0.54)	0.054	0.055	0.099
	Waaijenborg et al. [2008]	0.108	0.242	0.101	0.233
	Parkhomenko et al. [2009]	0.108	0.237	0.108	0.237
	Canonical ridge	0.128	0.276	0.128	0.276
	CCA	0.127	0.270	0.127	0.270
Noisy	SAR	0.067	0.158	0.113	0.242
	Witten and Tibshirani [2009]	0.058 (0.08)	0.179	0.116 (0.57)	0.224
	Waaijenborg et al. [2008]	0.179	0.370	0.175	0.363
	Parkhomenko et al. [2009]	0.192	0.386	0.192	0.386
	Canonical ridge	0.216	0.421	0.216	0.421
	CCA	0.211	0.425	0.211	0.425
Correlated	SAR	0.002	0.065	0.011	0.023
	Witten and Tibshirani [2009]	0.068	0.314	0.146	0.325
	Waaijenborg et al. [2008]	0.251	0.533	0.241	0.523
	Parkhomenko et al. [2009]	0.326	0.718	0.326	0.718
	Canonical ridge	0.049	0.044	0.049	0.044
	CCA	0.043	0.033	0.043	0.033
Sparse	SAR	0.139	0.244	0.535	0.619
High-dimensional	Witten and Tibshirani [2009]	0.261	0.394	0.448	0.476
	Waaijenborg et al. [2008]	0.854	0.961	0.845	0.958
	Parkhomenko et al. [2009]	0.826	0.924	0.826	0.924
	Canonical ridge	0.914	1.025	0.914	1.025
	CCA	1.086	1.198	1.086	1.198
NonSparse	SAR	0.370	0.356	0.422	0.425
High-dimensional	Witten and Tibshirani [2009]	0.833	0.846	0.961	0.932
	Waaijenborg et al. [2008]	1.284	1.306	1.253	1.277
	Parkhomenko et al. [2009]	1.147	1.166	1.147	1.166
	Canonical ridge	0.928	0.991	0.928	0.991
	CCA	1.291	1.345	1.291	1.345
UltraHigh-dimensional	SAR	1.402	1.370	1.399	1.376
	Witten and Tibshirani [2009]	1.546	1.523	1.546	1.539
	Waaijenborg et al. [2008]	1.546	1.536	1.546	1.536
	Parkhomenko et al. [2009]	1.528	1.519	1.528	1.519
	Canonical ridge	1.547	1.537	1.547	1.537

in the Correlated design.

Next, we compare the SAR algorithm to its sparse alternatives. In the Uncorrelated and Noisy design, the uncorrelatedness restriction imposed by Waaijenborg et al. [2008], Parkhomenko et al. [2009] and Witten and Tibshirani [2009] is satisfied. Therefore, we expect these methods to perform especially well. Nevertheless, even in this setting, the SAR algorithm performs competitive to the method of Witten and Tibshirani [2009] and significantly better than the other two. In the Correlated and High-dimensional designs this uncorrelatedness restriction is violated. Here, we see even more clearly that the SAR algorithm has a significant advantage over its sparse alternatives. In the correlated design, for instance, the SAR algorithm outperforms the method of Witten and Tibshirani [2009] by more than a factor 10 for the first canonical vector (i.e. estimation accuracy of 0.002 against 0.068), and by a factor 5 for the second canonical vector (i.e. estimation accuracy of 0.065 against 0.314). The gains in estimation accuracy of the SAR algorithm compared to the other sparse CCA methods are even more outspoken.

Table 5.3 compares the results on sparsity recognition performance for the sparse simulation designs among the sparse CCA techniques. The methods of Parkhomenko et al. [2009] and Waaijenborg et al. [2008] produce the least sparse solution, indicated by the high true positive rates and low true negative rates. The SAR algorithm and the method of Witten and Tibshirani [2009] tend to produce the most sparse solutions, indicated by the high true negative rates and low true positive rates. Contrary to sparse CCA, traditional CCA and the canonical ridge do not perform variable selection simultaneously with model estimation. Therefore, traditional CCA and canonical ridge are not included in Table 5.3. All elements of the canonical vectors are estimated as non-zero, resulting in a perfect true positive rate and zero true negative rate.

For fair comparison between the methods, we also compare the performance of the different methods when the tuning parameters are consistently selected (by maximizing test sample correlation using 5-fold cross-validation). Results are in the last columns of Table 5.2 and 5.3. For the method of Parkhomenko et al. [2009] and the canonical ridge, this cross-validation procedure was already suggested by the respective authors. In all designs (except the Sparse High-Dimensional design) the SAR algorithm remains the best performing. The relative performance of the CCA methods in terms of estimation accuracy remains unchanged when either cross-validation is used to select the sparsity parameters or the approach proposed by the respective authors. Note, however, that especially the SAR

Table 5.3: *Sparsity recognition performance: true positive rate and true negative rate for canonical vectors in the* **A** *and* **B** *matrices.*

Design	Method	Tuning parameters selected using							
		Author's method				Cross-validation			
		A		**B**		**A**		**B**	
		TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Uncorrelated	SAR	0.79	0.79	0.78	0.85	0.94	0.27	0.93	0.30
	Witten and Tibshirani [2009]	0.76	0.84	0.78	0.78	0.85	0.52	0.82	0.65
	Waaijenborg et al. [2008]	0.98	0.11	0.98	0.13	0.97	0.15	0.97	0.17
	Parkhomenko et al. [2009]	0.93	0.22	0.93	0.25	0.93	0.22	0.93	0.25
Noisy	SAR	0.69	0.61	0.66	0.70	0.93	0.19	0.92	0.21
	Witten and Tibshirani [2009]	0.64	0.66	0.67	0.62	0.77	0.41	0.70	0.54
	Waaijenborg et al. [2008]	0.96	0.10	0.95	0.11	0.93	0.14	0.94	0.14
	Parkhomenko et al. [2009]	0.90	0.15	0.91	0.16	0.90	0.15	0.91	0.16
Correlated	SAR	0.80	0.92	0.55	0.93	1.00	0.35	1.00	0.23
	Witten and Tibshirani [2009]	0.51	0.79	0.43	0.75	0.64	0.56	0.43	0.75
	Waaijenborg et al. [2008]	0.96	0.14	0.93	0.16	0.95	0.17	0.92	0.19
	Parkhomenko et al. [2009]	0.86	0.22	0.84	0.24	0.86	0.22	0.84	0.24
Sparse High-dimensional	SAR	0.40	0.93	0.34	0.94	0.82	0.43	0.70	0.49
	Witten and Tibshirani [2009]	0.39	0.85	0.32	0.84	0.44	0.74	0.31	0.85
	Waaijenborg et al. [2008]	0.86	0.25	0.83	0.27	0.88	0.23	0.84	0.25
	Parkhomenko et al. [2009]	0.74	0.35	0.70	0.38	0.74	0.35	0.70	0.38
UltraHigh-dimensional	SAR	0.20	0.89	0.22	0.89	0.35	0.85	0.36	0.86
	Witten and Tibshirani [2009]	0.18	0.84	0.19	0.84	0.29	0.72	0.31	0.72
	Waaijenborg et al. [2008]	0.95	0.06	0.94	0.06	0.94	0.06	0.95	0.06
	Parkhomenko et al. [2009]	0.36	0.67	0.26	0.79	0.36	0.67	0.26	0.79

algorithm and the method of Witten and Tibshirani [2009] perform better in terms of estimation accuracy when following the respective authors' proposal to select the sparsity parameters (except in the UltraHigh-Dimensional design). Looking at sparsity recognition performance, the SAR algorithm and method of Witten and Tibshirani [2009] now show higher values of true positive rate and lower values of true negative rate.

Convergence properties of the SAR algorithm - using BIC or $5-$fold cross-validation (CV) to select the sparsity parameters - are reported in Table 5.4. We report the average (averaged across simulation runs) number of iterations up to convergence and the percentage of non-convergence (i.e. percentage of simulation runs where convergence was not reached after 100 iterations). In the simulations we conducted, the SAR algorithm almost always reached convergence.[1]

To conclude, as we can see from Table 5.2, overall, the SAR algorithm did perform significantly better than the other sparse CCA methods. The advantage of our sparse CCA approach over the other sparse CCA approaches is most

[1] In case of non-convergence, results from the last iteration run are taken.

Table 5.4: *Convergence properties of the SAR algorithm. Results are reported when using BIC or 5−fold cross-validation to select the sparsity parameter.*

Design	Average number of iterations		Percentage of non-convergence	
	BIC	CV	BIC	CV
Uncorrelated	3.67	4.55	0.0%	0.0%
Correlated	5.11	4.29	0.0%	0.0%
Noisy	4.17	5.52	0.0%	0.1%
Sparse High-dimensional	7.98	11.42	0.0%	0.0%
NonSparse High-dimensional	5.83	11.00	1.0%	0.6%
UltraHigh-dimensional	20.63	17.17	0.0%	0.0%

outspoken when the data sets contain correlated variables.

5.5 Genomic data application

In recent years, high-dimensional genomic data sets have arisen, containing thousands of gene expression and other phenotype measurements (e.g., Hommel and Kropf, 2005, Lauter et al., 2009). We use the publicly available breast cancer data set described in Chin et al. [2006] and available in the `R package PMA` [Witten et al., 2011]. Comparative genomic hybridization (CGH) data (2149 variables) and gene expression data (19 672 variables) are available on 89 samples. The objective is to identify copy number change variables that are correlated with a subset of gene expression variables. Copy number changes on a particular chromosome are associated with expression changes in genes located on the same chromosome [Witten and Tibshirani, 2009]. Therefore, we analyze the data for each chromosome separately, each time using the CGH and gene expression variables for that particular chromosome. The dimension of both sets of variables is large compared to the sample size such that traditional CCA cannot be performed. In such a high-dimensional setting, the use of sparse CCA techniques is appealing. We use the SAR algorithm to perform sparse CCA for each chromosome separately.

To decide on the number of canonical variates pairs to extract, we apply the canonical ridge to each chromosome. Figure 5.1 shows the first 20 estimated canonical correlations for each of the 23 chromosomes. For each chromosome, we use the maximum eigenvalue ratio criterion, discussed in Section 5.3, to determine the number of canonical variate pairs to extract. Depending on the specific chromosome, this criterion indicates to extract either 1, 2, 3 or 4 canonical variate pairs.

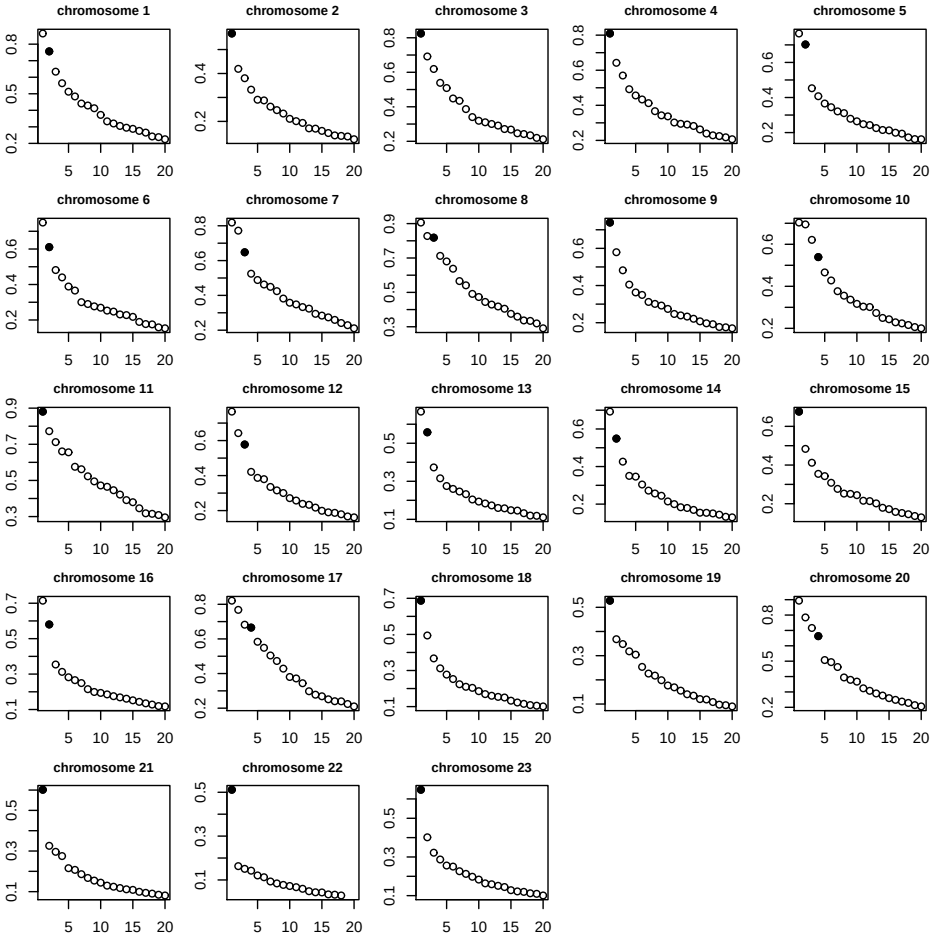


Figure 5.1: *Estimated canonical correlations using the canonical ridge, for each of the 23 chromosomes. The highest order pair of canonical variates to retain, as selected by the maximum eigenvalue ratio criterion, is indicated by a solid black circle.*

To compare the performance of the SAR algorithm to the other sparse CCA procedures discussed in Section 5.4, we perform an out-of-sample cross-validation exercise.[2] This out-of-sample exercise enables an evaluation of the performance of the different methods without knowing the distribution of the data. More precisely, for each chromosome, we perform a leave-one-out cross-validation exercise and compute the cross-validation score

$$CV = \frac{1}{n}\sum_{i=1}^{n} ||\widehat{\mathbf{A}}_{-i}^{T}\mathbf{x}_i - \widehat{\mathbf{B}}_{-i}^{T}\mathbf{y}_i||^2,$$

where $\widehat{\mathbf{A}}_{-i}^{T}$ and $\widehat{\mathbf{B}}_{-i}^{T}$ contain the estimated canonical vectors when the i^{th} observation is left out of the estimation sample. For each chromosome, the number of estimated canonical vector pairs used in the calculation of the cross-validation score differs and corresponds to the number determined by the maximum eigenvalue ratio criterion (see Figure 5.1). We use leave-one-out cross-validation, which corresponds to n-fold cross-validation, to have sufficient data points on which we can compare the performance of the different methods. We compute this cross-validation score for each of the sparse CCA techniques. The technique that leads to the lowest value of this cross-validation score achieves the best out-of-sample performance.

Averaged across all chromosomes, the SAR algorithm attains a cross-validation score of 104.00, the method of Witten and Tibshirani [2009] 223.08, Parkhomenko et al. [2009] 2778.57 and Waaijenborg et al. [2008] 680.05. Thus, the SAR algorithm outperforms its alternatives. Furthermore, we compute relative cross-validation scores, being the cross-validation score of a method relative to the cross-validation score of the SAR algorithm. The relative cross-validations scores on a logarithmic scale (23 scores, one for each chromosome) are presented in Figure 5.2. A value of the relative cross-validation score larger than 1 (horizontal dashed line) indicates better performance of the SAR algorithm. The SAR algorithm always attains the best cross-validation score, except for three cases out of 23 where the methods of Witten and Tibshirani [2009] and Waaijenborg et al. [2008] achieve a lower cross-validation score. The differences in performance compared to the method of Parkhomenko et al. [2009] and Waaijenborg et al. [2008] are large. The cross-validation scores obtained with the SAR algorithm and the method of Witten and Tibshirani [2009] are substantially lower than those obtained with the method of Parkhomenko et al. [2009] and Waaijenborg et al. [2008]. The solutions obtained with the former two are much sparser than

[2] For all methods, tuning parameters are selected according to the authors' method.

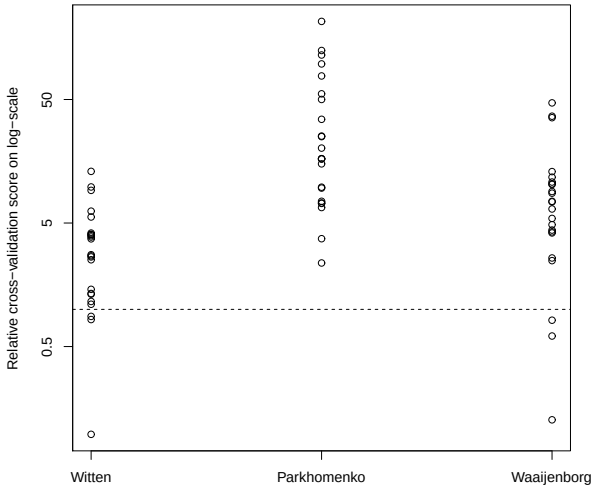


Figure 5.2: *Cross-validation scores on logarithmic scale (23, one for each chromosome) of Witten and Tibshirani [2009], Parkhomenko et al. [2009] and Waaijenborg et al., relative to the SAR algorithm. The horizontal dashed line at 1 indicates the relative cross-validation score of the SAR algorithm.*

the ones obtained with the latter two. Sparsity thus helps in achieving a good cross-validation score.

The dependency structure within each set of variables might explain the good performance of the SAR algorithm relative to its alternatives. For the first chromosome, for instance, 20% of the (absolute) correlations between the 136 CGH spots are larger than 0.6. The same holds for the other chromosomes. In the simulation study from Section 5.4, we show that the SAR algorithm performs much better for highly correlated data sets than the other sparse CCA techniques, that impose prior uncorrelatedness restrictions. This might explain why the SAR algorithm outperforms its alternatives in the out-of-sample cross-validation exercise.

Next, we discuss the solution provided by the SAR algorithm. For each chromosome, sparse canonical vectors are obtained. We do not fix the number of non-zero elements in the canonical vectors in advance, but select the sparsity parameter using the BIC discussed in Section 5.3. Figure 5.3 represents for each

chromosome the copy number change measurements with non-zero weights. The construction of this figure is similar to the one presented in Witten and Tibshirani [2009]. We use the `R-code` available in the `R package PMA` [Witten et al., 2011]. Each CGH spot has a certain position on a chromosome, called the nucleotide position. The CGH measurements selected by the SAR algorithm are indicated by plotting a vertical line on their respective nucleotide position. The four panels indicate the subset of variables selected in the construction of the corresponding canonical variate pair (first pair: top left, second pair: top right, third pair: bottom left, fourth pair: bottom right).

We see from Figure 5.3 that the degree of sparsity selected by the BIC varies from one chromosome to the other. For chromosome 23, for example, only one canonical variate pair is selected and the BIC suggests a very sparse canonical vector. For chromosome 9, also one canonical variate pair is extracted but with a larger number of non-zero elements. However, a lot of non-zero weights are small in magnitude which can be seen from the length of the vertical lines. By adjusting the sparsity parameter to a higher value, a sparser solution could be obtained. A trade-off needs to be made between inducing more sparsity and thus performing better noise filtering, on the one hand, and reducing the risk of not including all important variables, on the other hand. Depending on the researcher's objective, the desired level of sparsity can be easily controlled by adjusting the sparsity parameter.

5.6 Conclusion

In high-dimensional settings, the estimation imprecision of traditional CCA will be large. An appropriate choice of the sample size is key to tackle estimation imprecision. However, genomewide association studies are often constrained by cost, and sample sizes are often limited by clinical samples that are well characterized [Spencer et al., 2009]. In such studies, penalized CCA methods play an important role. The canonical ridge still includes all variables in the canonical vectors, whereas sparse CCA only includes a subset of the variables. This is highly valuable in high-dimensional settings since it eases interpretation, as illustrated in the genomic data application.

In this paper, we introduce a Sparse Alternating Regression (SAR) algorithm that considers the CCA problem from a predictive point of view. We recast the CCA problem into a penalized alternating regression framework to obtain sparse canonical vectors. Contrary to other popular sparse CCA procedures (i.e. Witten

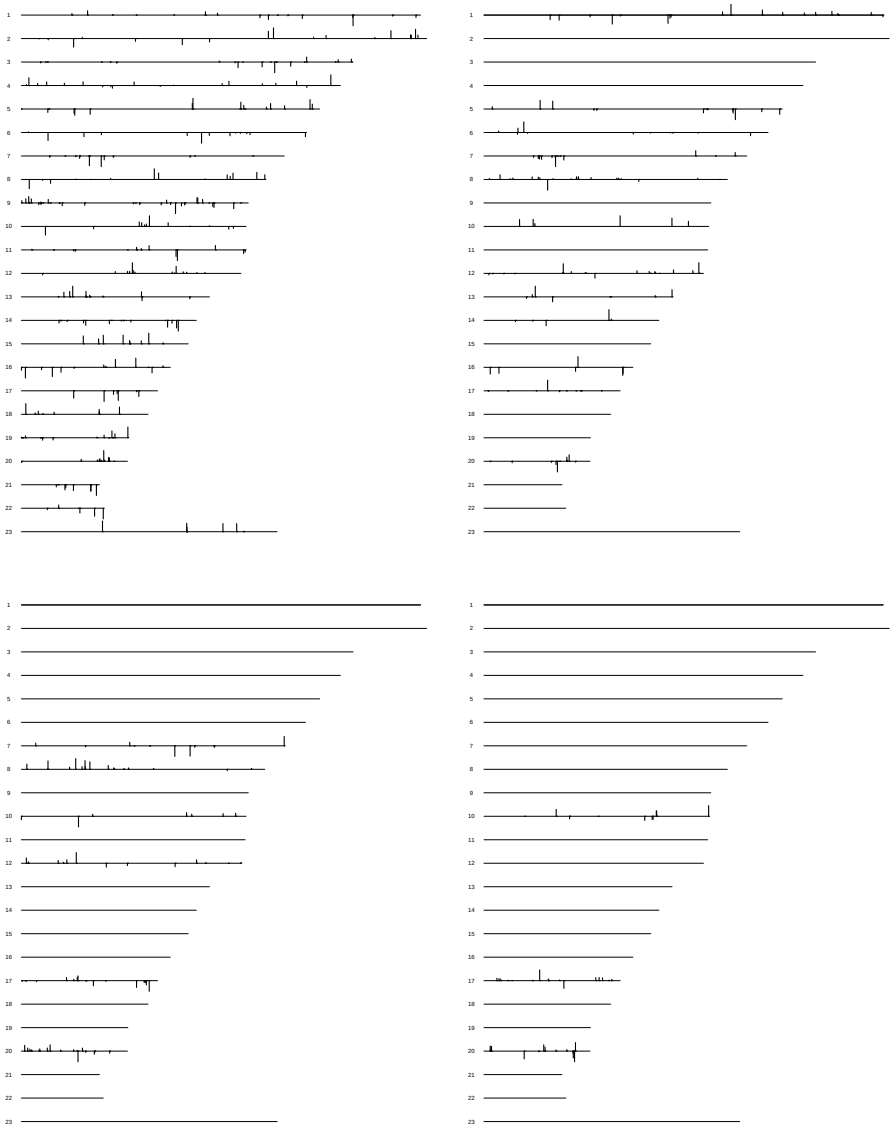


Figure 5.3: *SAR algorithm: copy number change measurements with non-zero weights in the first (top left), the second (top right), the third (bottom left) and the fourth (bottom right) canonical vectors are indicated for each of the 23 chromosomes.*

and Tibshirani, 2009; Parkhomenko et al., 2009; Waaijenborg et al., 2008), we do not impose any restriction on the correlation between variables. This leads to an important advantage of the SAR algorithm compared to the competing sparse CCA methods. Indeed, we show that the SAR algorithm produces much better results than the other sparse CCA approaches. Especially in simulation settings when there is a dependency structure within each set of variables, the gains in estimation accuracy achieved by the SAR algorithm are outspoken. Also in the genomic data application, the data sets contain highly correlated variables. We illustrate that the SAR algorithm considerably outperforms the other sparse CCA techniques in an out-of-sample cross-validation exercise.

Both the SAR algorithm and the method of Waaijenborg et al. [2008] use an alternating regression framework. There are, however, two important differences between both approaches, leading towards significant differences in performance. First, Waaijenborg et al. [2008] perform univariate soft thresholding, which ignores the dependency structure within each set of variables. In contrast, we apply the lasso penalty to multiple linear regressions. The lasso only equals the soft thresholding estimator for a linear model with orthonormal design (see e.g. Donoho and Johnstone, 1994). Secondly, we express the higher order canonical vectors in terms of the original data sets, whereas Waaijenborg et al. [2008] express them in terms of the deflated data matrices. An unreported simulation study indicates that especially the first difference leads to important differences in performance.

In this paper, a lasso penalty is used to induce sparsity. Future work might consider other choices of penalty functions (see Prabhakar and Fridley, 2012). For instance, the adaptive lasso [Zou, 2006], the smoothly clipped absolute deviation (SCAD) penalty [Fan and Li, 2001], or a lasso with positivity constraints (see Lykou and Whittaker, 2010). Note that Lykou and Whittaker [2010] also treat CCA as a least squares problem. They focus on orthogonality properties of CCA and only construct the first two pairs of sparse canonical vectors. Their approach could be extended to higher order canonical correlations, but this would increase the number of orthogonality constraints and the computing time substantially. The level of sparsity produced by all sparse CCA techniques hinges on the selection method used for the sparsity parameters. This might lead to substantial differences in sparsity recognition performance, as illustrated in the simulation study. Future work still needs to be done on the comparison of methods (BIC, cross-validation, measure of explained variability, among others) to select the optimal value of the tuning parameters.

Chapter 6

Robust sparse canonical correlation analysis

Abstract

Canonical correlation analysis (CCA) is a multivariate statistical method which describes the associations between two sets of variables. The objective is to find linear combinations of the variables in each data set having maximal correlation. This paper discusses a method for Robust Sparse CCA. Sparse estimation produces canonical vectors with some of their elements estimated as exactly zero. As such, their interpretability is improved. Sparse estimation can also be used to analyze high-dimensional data sets that are often found in the field of biometrics. Robust methods can cope with outliers in the data that are likely to occur in high-dimensional data sets. We illustrate the good performance of the Robust Sparse CCA method by several simulation studies and three biometric examples. Robust Sparse CCA performs much better than other CCA methods.

6.1 Introduction

Canonical correlation analysis (CCA), introduced by Hotelling [1936], identifies and quantifies the associations between two sets of variables. CCA searches for linear combinations, called *canonical variates*, of each of the two sets of variables having maximal correlation. The coefficients of these linear combinations are called the *canonical vectors*. The correlations between the canonical variates are called the *canonical correlations*. CCA is used to study associations in, for

instance, genomic data [van Wieringen and van de Wiel, 2009], environmental data [Iaci et al., 2010], or biomedical data [Alonso et al., 2003]. For more information on canonical correlations analysis, see e.g. Johnson and Wichern (1998, Chapter 10).

Sparse canonical vectors are canonical vectors with some of their elements estimated as exactly zero. The canonical variates then only depend on a subset of the variables, those corresponding to the non-zero elements of the estimated canonical vectors. Hence, the canonical variates are easier to interpret, in particular for high-dimensional data sets. Examples of CCA for high-dimensional data sets can be found in, for example, genetics [Gonzalez et al., 2008, Prabhakar and Fridley, 2012, Cruz-Cano and Lee, 2014] and machine learning [Sun et al., 2011].

Different approaches for sparse CCA have been proposed in the literature. Parkhomenko et al. [2009] use a sparse singular value decomposition to derive sparse singular vectors. Witten and Tibshirani [2009] develop a penalized matrix decomposition, and show how to apply it for sparse CCA. Waaijenborg et al. [2008], Lykou and Whittaker [2010], An et al. [2013] and Wilms and Croux [2015] convert the CCA problem into a penalized regression framework to produce sparse canonical vectors. All these methods are not robust to outliers. A common problem in multivariate data sets, however, is the frequent occurrence of outliers. In genomics, for instance, some patients can react very differently to treatments because of their individual-specific genetic structure. Therefore, the possible presence of outlying observations should be taken into account.

Several *robust CCA* methods have been introduced in the literature. Dehon and Croux [2002] considers robust CCA using the Minimum Covariance Determinant (MCD, Rousseeuw and Van Driessen, 1999) estimator. Asymptotic properties for CCA based on robust estimators of the covariance matrix are discussed in Taskinen et al. [2006]. Branco et al. [2005] use a robust alternating regression approach to obtain the canonical variates. CCA can also be considered as a prediction problem, where the canonical variates obtained from the first data set serve as optimal predictors for the canonical variates of the second data set, and vice versa. As such, Adrover and Donato [2015] use a robust M-scale to evaluate the prediction quality, whereas the approach of Kudraszow and Maronna [2011] is based on a robust estimator for the multivariate linear model. None of these methods, however, are sparse.

This paper proposes a CCA method that is sparse and robust at the same time. As such, we deal with two important topics in applied statistics: sparse model estimation and the presence of outliers in the data. We use an alternating

robust, sparse regression framework to sequentially obtain the canonical variates. We obtain sparse canonical vectors that are resistant to outlying observations by using the sparse Least Trimmed Squares (sparse LTS) estimator of Alfons et al. [2013]. Robust Sparse CCA has clear advantages: (i) Robust Sparse CCA provides well interpretable canonical vectors since some of the elements of the canonical vectors are estimated as exactly zero, (ii) Robust Sparse CCA is still computable for high-dimensional data sets, where the sample size exceeds the number of variables in each data set, and (iii) Robust Sparse CCA can cope with outliers in the data, which are even more likely to occur in high dimensions.

The remainder of this article is organized as follows. Section 6.2 considers the robust and sparse estimator for the CCA problem. Section 6.3 discusses the algorithm. Section 6.4 presents simulation results where we compare Robust Sparse CCA to standard CCA, Robust CCA and Sparse CCA. In Section 6.5, we show that Robust Sparse CCA performs much better than the other methods on three biometric data sets. Section 6.6 concludes.

6.2 The estimator

We consider the CCA problem in a regression framework, as proposed by Brillinger [1975] and Izenman [1975]. Given a sample of n observations $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{y}_i \in \mathbb{R}^q$ $(i = 1, \ldots, n)$. The two data matrices are denoted as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$. We assume the data matrices are robustly centered using the median. The estimated canonical vectors are collected in the columns of the matrices $\widehat{\mathbf{A}} \in \mathbb{R}^{p \times r}$ and $\widehat{\mathbf{B}} \in \mathbb{R}^{q \times r}$. Here r is the number of canonical vectors. The columns of the matrices $\mathbf{X}\widehat{\mathbf{A}}$ and $\mathbf{Y}\widehat{\mathbf{B}}$ contain the estimates of the realizations of the canonical variates, and we denote their j^{th} column by $\hat{\mathbf{u}}_j$ and $\hat{\mathbf{v}}_j$, for $1 \leq j \leq r$. The objective function defining the canonical vector estimates is

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \underset{(\mathbf{A},\mathbf{B})}{\operatorname{argmin}} \sum_{i=1}^{n} ||\mathbf{A}^T x_i - \mathbf{B}^T \mathbf{y}_i||^2. \tag{6.1}$$

The objective function in (6.1) is minimized under the restriction that each canonical variate $\hat{\mathbf{u}}_j$ is uncorrelated with the lower order canonical variates $\hat{\mathbf{u}}_k$, with $1 \leq k < j \leq r$. Similarly for the canonical vectors within the second set of variables. For identification purpose, a normalization condition requiring the canonical vectors to have unit norm is added. Typically, the canonical vectors are obtained by an eigenvalue analysis of a certain matrix involving the inverses of sample covariance matrices. But if $n < \max(q, p)$, these inverses do not exist.

We estimate the canonical vectors with an alternating regression procedure. If the matrix $\mathbf{A}$ in (6.1) is kept fixed, the matrix $\mathbf{B}$ can be obtained from a Least Squares regression of the canonical variates on $\mathbf{y}$ (and vice versa for estimating $\mathbf{A}$ keeping $\mathbf{B}$ fixed). The standard Least Squares estimator, however, is not sparse, nor robust to outliers. Therefore, we replace it by the sparse Least Trimmed Squares (sparse LTS) estimator [Alfons et al., 2013]. The sparse LTS estimator can be applied to high-dimensional data and is robust to outliers.

6.3 The algorithm

We use a sequential algorithm to derive the canonical vectors.

First canonical vector pair. Denote the first canonical vector pair by $(\mathbf{A}_1, \mathbf{B}_1)$. Assume that the value of $\mathbf{A}_1$ is known. Denote the vector of squared residuals by $\mathbf{r}^2(\mathbf{B}_1) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\mathbf{A}_1^T\mathbf{x}_i - \mathbf{B}_1^T\mathbf{y}_i)^2, i = 1, \ldots, n$. The estimate of $\mathbf{B}_1$ is obtained as

$$\widehat{\mathbf{B}}_1|\mathbf{A}_1 = \underset{\mathbf{B}_1}{\operatorname{argmin}} \sum_{i=1}^{h} \left(\mathbf{r}^2(\mathbf{B}_1)\right)_{i:n} + h\lambda_{B_1} \sum_{j=1}^{q} |b_{1j}|, \tag{6.2}$$

where $\lambda_{B_1} > 0$ is a sparsity parameter, b_{1j} is the j^{th} element, $j = 1, \ldots, q$, of the first canonical vector $\mathbf{B}_1$, and $\left(\mathbf{r}^2(\mathbf{B}_1)\right)_{1:n} \leq \ldots \leq \left(\mathbf{r}^2(\mathbf{B}_1)\right)_{n:n}$ are the order statistics of the squared residuals. The canonical vector $\widehat{\mathbf{B}}_1$ is normed to length 1. The solution to (6.2) equals the sparse LTS estimator with $\mathbf{XA}_1$ as response and $\mathbf{Y}$ as predictor. Regularization by adding a penalty term to the objective function is necessary since the design matrix $\mathbf{Y}$ can be high-dimensional. Sparse model estimates are obtained by adding an L_1 penalty to the LTS objective function, similar as for the lasso regression estimator [Tibshirani, 1996]. The sparse LTS estimator is computed with trimming proportion 25%, so size of the subsample $h = \lfloor 0.75n \rfloor$. To increase efficiency, we use a reweighting step.[1] As such, we get a robust sparse estimate $\widehat{\mathbf{B}}_1$.

Analogously, for a fixed value $\mathbf{B}_1$, denote the vector of squared residuals by $\mathbf{r}^2(\mathbf{A}_1) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\mathbf{B}_1^T\mathbf{y}_i - \mathbf{A}_1^T\mathbf{x}_i)^2, i = 1, \ldots, n$. The sparse LTS

[1] The reweighted sparse LTS is the lasso estimator computed from the observations not detected as outliers by the sparse LTS, i.e. having an absolute value of the standardized residuals smaller than or equal to the 98.75^{th} quantile of the standard normal distribution (see Alfons et al., 2013 for more detail).

regression estimate of $\mathbf{A}_1$ with $\mathbf{YB}_1$ as response and $\mathbf{X}$ as predictor is given by

$$\widehat{\mathbf{A}}_1|\mathbf{B}_1 = \underset{\mathbf{A}_1}{\operatorname{argmin}} \sum_{i=1}^{h} \left(\mathbf{r}^2(\mathbf{A}_1)\right)_{i:n} + h\lambda_{A_1} \sum_{j=1}^{p} |a_{1j}|, \tag{6.3}$$

where $\lambda_{A_1} > 0$ is a sparsity parameter, a_{1j} is the j^{th} element, $j = 1, \ldots, p$ of the first canonical vector $\mathbf{A}_1$, and $\left(\mathbf{r}^2(\mathbf{A}_1)\right)_{1:n} \leq \ldots \leq \left(\mathbf{r}^2(\mathbf{A}_1)\right)_{n:n}$ are the order statistics of the squared residuals. The canonical vector $\widehat{\mathbf{A}}_1$ is normed to length 1.

This leads to an alternating regression scheme, updating in each step the estimates of the canonical vectors until convergence.

Higher order canonical vector pairs. We use deflated data matrices to estimate the higher order canonical vector pairs (see e.g. Branco et al., 2005). For the second canonical vector pair, the deflated matrices are $\mathbf{X}_2^*$, the residuals of a column-by-column LTS regression of $\mathbf{X}$ on all lower order canonical variates, $\hat{\mathbf{u}}_1$ in this case; and $\mathbf{Y}_2^*$, the residuals of a column-by-column LTS regression of $\mathbf{Y}$ on $\hat{\mathbf{v}}_1$. Since these regressions only involve a small number of regressors, the standard LTS estimator with $\lambda = 0$ can be used.

The second canonical variate pair is then obtained by alternating between the following regressions until convergence:

$$\widehat{\mathbf{B}}_2^*|\mathbf{A}_2^* = \underset{\mathbf{B}_2^*}{\operatorname{argmin}} \sum_{i=1}^{h} \left(\mathbf{r}^2(\mathbf{B}_2^*)\right)_{i:n} + h\lambda_{B_2^*} \sum_{j=1}^{q} |b_{2j}^*|, \tag{6.4}$$

where $\mathbf{r}^2(\mathbf{B}_2^\star) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\mathbf{A}_2^{*T}\mathbf{x}_{2,i}^\star - \mathbf{B}_2^{\star T}\mathbf{y}_{2,i}^\star)^2, i = 1, \ldots, n$.

$$\widehat{\mathbf{A}}_2^*|\mathbf{B}_2^* = \underset{\mathbf{A}_2^*}{\operatorname{argmin}} \sum_{i=1}^{h} \left(\mathbf{r}^2(\mathbf{A}_2^*)\right)_{i:n} + h\lambda_{A_2^*} \sum_{j=1}^{p} |a_{2j}^*|, \tag{6.5}$$

where $\mathbf{r}^2(\mathbf{A}_2^\star) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\mathbf{B}_2^{*T}\mathbf{y}_{2,i}^\star - \mathbf{A}_2^{\star T}\mathbf{x}_{2,i}^\star)^2, i = 1, \ldots, n$. The canonical vectors $\widehat{\mathbf{B}}_2^*$ and $\widehat{\mathbf{A}}_2^*$ are both normed to length 1. We obtain $\hat{\mathbf{u}}_2^* = \mathbf{X}_2^*\widehat{\mathbf{A}}_2^*$ and $\hat{\mathbf{v}}_2^* = \mathbf{Y}_2^*\widehat{\mathbf{B}}_2^*$.

Finally, the second canonical vector needs to be expressed as linear combinations of the columns of the original data matrices, and not the deflated ones. Since we want to allow for zero coefficients in these linear combinations, a sparse approach is needed. To obtain a sparse $\widehat{\mathbf{A}}_2$, we regress $\hat{\mathbf{u}}_2^*$ on $\mathbf{X}$ using the sparse LTS estimator, yielding the fitted values $\hat{\mathbf{u}}_2 = \mathbf{X}\widehat{\mathbf{A}}_2$. To obtain a sparse $\widehat{\mathbf{B}}_2$, we regress $\hat{\mathbf{v}}_2^*$ on $\mathbf{Y}$ using the sparse LTS estimator, yielding the fitted values $\hat{\mathbf{v}}_2 = \mathbf{Y}\widehat{\mathbf{B}}_2$.

The higher order canonical variate pairs are obtained in a similar way. We perform alternating sparse LTS regressions as in (6.4) and (6.5), followed by a final sparse LTS step to retrieve the estimated canonical vectors $(\widehat{\mathbf{A}}_k, \widehat{\mathbf{B}}_k)$. It is not really necessary to use a sparse approach in regressions (6.4) and (6.5), other penalty functions can be used.

Initial value. A starting value for $\mathbf{A}_1$ is required to start up the algorithm. We compute the first robust principal component of $\mathbf{Y}$, denoted $\mathbf{z}_1$. The first robust principal component is calculated from the first eigenvector of the robustly estimated covariance matrix. For this aim, we use the spatial sign covariance estimator [Visuri et al., 2000]. We regress $\mathbf{z}_1$ on $\mathbf{X}$ using the sparse LTS. The estimated regression coefficient matrix of this regression is used as initial value for $\mathbf{A}_1$. To obtain an initial estimate for the higher order canonical vectors $\mathbf{A}_l$, for $l = 2, \ldots, r$, we use the first robust principal component of the deflated data matrix and proceed analogously.

Number of canonical variates to extract. To decide on the number of canonical variates r to extract, we use the maximum eigenvalue ratio criterion of An et al. [2013]. We apply the Robust Sparse CCA algorithm and calculate the robust correlations $\hat{\rho}_1, \ldots, \hat{\rho}_{\text{rmax}}$, with $\text{rmax} = \min(p, q, 10)$. For high-dimensional data sets, we consider a maximum of 10 canonical correlations, since in practice, more than 10 canonical vector pairs are never used. Each $\hat{\rho}_j$ is obtained by computing the correlation between $\hat{\mathbf{v}}_j$ and $\hat{\mathbf{u}}_j$ from the bivariate Minimum Covariance Determinant estimator with 25% trimming. Let $\hat{k}_j = \hat{\rho}_j / \hat{\rho}_{j+1}$ for $j = 1, \ldots, \text{rmax} - 1$. We extract r pairs of canonical variates, where $r = \text{argmax}_j \hat{k}_j$.

Convergence criterion. In each step of the alternating regression algorithm we update the estimates of the canonical vectors $\widehat{\mathbf{B}}_l^*$ and $\widehat{\mathbf{A}}_l^*$, for $l = 1, \ldots, r$. We iterate until the relative change in the value of the convergence criterion in two successive iterations is smaller than the convergence tolerance value $\epsilon = 10^{-2}$. As convergence criterion, we consider

$$\text{Convergence criterion} = \frac{1}{h} \sum_{i=1}^{h} (\mathbf{r}^2(\widehat{\mathbf{A}}_l^*, \widehat{\mathbf{B}}_l^*))_{i:n},$$

for $l = 1, \ldots, r$, where $\mathbf{r}^2(\widehat{\mathbf{A}}_l^*, \widehat{\mathbf{B}}_l^*) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\widehat{\mathbf{A}}_l^{*T} \mathbf{x}_{l,i}^\star - \widehat{\mathbf{B}}_l^{*T} \mathbf{y}_{l,i}^\star)^2$, $i = 1, \ldots, n$. $\mathbf{X}_l^*$ and $\mathbf{Y}_l^*$ are the original data sets for $l = 1$, and the deflated data matrices for $l = 2, \ldots, r$. In the simulations we conducted, convergence was almost always reached.

Choice of the sparsity parameter. The sparsity parameters controlling the penalization on the regression coefficient matrices are selected with the Bayesian Information Criterion (e.g. Yin and Li, 2011). We use a range of values for the sparsity parameters and select the one with the lowest value of

$$\mathrm{BIC}_{\lambda_{\widehat{\mathbf{A}}_l^*}} = n \cdot \log\left(\frac{1}{h}\sum_{i=1}^{h}\left(\mathbf{r}^2(\widehat{\mathbf{A}}_l^*)\right)_{i:n}\right) + df_{\lambda_{\widehat{\mathbf{A}}_l^*}} \cdot \log(n),$$

$$\mathrm{BIC}_{\lambda_{\widehat{\mathbf{B}}_l^*}} = n \cdot \log\left(\frac{1}{h}\sum_{i=1}^{h}\left(\mathbf{r}^2(\widehat{\mathbf{B}}_l^*)\right)_{i:n}\right) + df_{\lambda_{\widehat{\mathbf{B}}_l^*}} \cdot \log(n),$$

for $l = 1, \ldots, r$, with $df_{\lambda_{\widehat{\mathbf{A}}_l^*}}$ and $df_{\lambda_{\widehat{\mathbf{B}}_l^*}}$ the respective number of non-zero estimated regression coefficients.

6.4 Simulation Study

We compare the performance of the Robust Sparse CCA method with (i) standard CCA, (ii) Robust CCA, and (iii) Sparse CCA. The alternating regression algorithm is used for all 4 estimators, for ease of comparability. Robust CCA uses LTS instead of sparse LTS, and corresponds to the alternating regression approach of Branco et al. [2005]. Sparse CCA uses the lasso instead of sparse LTS, Pearson correlations for computing the canonical correlations, and ordinary PCA for getting the initial values. Standard CCA is like sparse CCA, but using the LS instead of the lasso.

6.4.1 Design

Several simulation designs are considered. In the first simulation design (revised from Branco et al., 2005), there is one canonical variate pair and the canonical vectors have a sparse structure. The canonical vectors are very sparse; each containing only one non-zero element. In the second design, there are two canonical variate pairs and the canonical vectors are non-sparse. In the third and fourth design, there are a lot of variables ($p = 100$) compared to the sample size ($n = 100$). In design three, there is one canonical variate pair and the canonical vectors are sparse. In design four, there are two canonical variate pairs and the canonical vectors are non-sparse. Only Sparse CCA and Robust Sparse CCA can be computed in design three and four. The number of simulations for each setting is $M = 1000$.

Table 6.1: *Simulation designs.*

Simulation design	n	p	q	$\mathbf{\Sigma}_{xx}$	$\mathbf{\Sigma}_{yy}$	$\mathbf{\Sigma}_{xy}$
Sparse Low-dimensional	100	6	4	$0.01\mathbf{I}_p$	$0.01\mathbf{I}_q$	$\begin{bmatrix} 0.009 & \mathbf{0}_{1\times 3} \\ \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 3} \end{bmatrix}$
NonSparse Low-dimensional	100	12	8	$0.01\mathbf{I}_p$	$0.01\mathbf{I}_q$	$\mathbf{0.001}_{p\times q}$
Sparse High-dimensional	100	100	4	$0.1\mathbf{I}_p$	$0.1\mathbf{I}_q$	$\begin{bmatrix} \mathbf{0.045}_{2\times 2} & \mathbf{0}_{2\times 2} \\ \mathbf{0}_{98\times 2} & \mathbf{0}_{98\times 2} \end{bmatrix}$
NonSparse High-dimensional	100	100	4	$0.1\mathbf{I}_p$	$0.1\mathbf{I}_q$	$\begin{bmatrix} \mathbf{0.045}_{2\times 2} & \mathbf{0.001}_{2\times 2} \\ \mathbf{0.001}_{98\times 2} & \mathbf{0.001}_{98\times 2} \end{bmatrix}$

For each design, the following settings are considered

(a) *No contamination.* We generate data matrices $\mathbf{X}$ and $\mathbf{Y}$ according to a multivariate normal distribution $N_{p+q}(\mathbf{0}, \mathbf{\Sigma})$, with covariance matrices described in Table 6.1.

(b) *t-distribution.* We generate data matrices $\mathbf{X}$ and $\mathbf{Y}$ according to a multivariate t-distribution with three degrees of freedom $t_3(\mathbf{0}, \mathbf{\Sigma})$.

(c) *Contamination.* 90% of the data are generated from $N_{p+q}(\mathbf{0}, \mathbf{\Sigma})$, and 10% of the data are generated from $N_{p+q}(\mathbf{2}, \mathbf{\Sigma}_{\text{cont}})$, with

$$\mathbf{\Sigma}_{\text{cont}} = \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{yy} \end{bmatrix}.$$

Similar conclusions can be drawn from other contamination settings (e.g. where only one of the two data sets is contaminated) and are available from the authors upon request.

6.4.2 Performance measures

The estimators are evaluated on their estimation accuracy. We compute for each simulation run m, with $m = 1, \ldots, M = 1000$, the angle $\theta^m(\hat{\mathbf{A}}^{(m)}, \mathbf{A})$ between the subspace spanned by the estimated canonical vectors (contained in the columns of $\hat{\mathbf{A}}^{(m)}$) and the subspace spanned by the true canonical vectors (contained in the columns of $\mathbf{A}$). We proceed analogously for the matrix $\mathbf{B}$. The average angles, measuring the estimation accuracy, are given by

$$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A}) = \frac{1}{M} \sum_{m=1}^{M} \theta^m(\hat{\mathbf{A}}^{(m)}, \mathbf{A}) \quad \text{and} \quad \bar{\theta}(\hat{\mathbf{B}}, \mathbf{B}) = \frac{1}{M} \sum_{m=1}^{M} \theta^m(\hat{\mathbf{B}}^{(m)}, \mathbf{B}).$$

For evaluating sparsity, we use the true positive rate and the true negative rate (e.g. Rothman et al., 2010)

$$\text{TPR}(\hat{\mathbf{A}}, \mathbf{A}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\#\{(i,j) : \widehat{\mathbf{A}}_{ij}^{(m)} \neq 0 \text{ and } \mathbf{A}_{ij} \neq 0\}}{\#\{(i,j) : \mathbf{A}_{ij} \neq 0\}}$$
$$\text{TNR}(\hat{\mathbf{A}}, \mathbf{A}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\#\{(i,j) : \widehat{\mathbf{A}}_{ij}^{(m)} = 0 \text{ and } \mathbf{A}_{ij} = 0\}}{\#\{(i,j) : \mathbf{A}_{ij} = 0\}}.$$

We proceed analogously for the matrix $\mathbf{B}$. A true positive is a coefficient that is non-zero in the true model, and is estimated as non-zero. A true negative is a coefficient that is zero in the true model, and is estimated as zero. Both the true positive rate and the true negative rate should be as high as possible for a sparse estimator.

6.4.3 Results

Summary results for the estimator $\widehat{\mathbf{A}}$ are in Table 6.2. The results for the estimator $\widehat{\mathbf{B}}$ are similar and are, therefore, omitted.

First we discuss the results from the Sparse Low-dimensional design. In the scenario without contamination, the sparse estimators Sparse CCA and Robust Sparse CCA achieve a much better average estimation accuracy than the non-sparse estimators CCA and Robust CCA. As expected, a sparse method results in increased estimation accuracy when the true canonical vectors have a sparse structure. Looking at sparsity recognition performance, Sparse CCA and Robust Sparse CCA perform equally good in retrieving the sparsity in the data generating process. In the contaminated simulation setting, the robust estimators maintain their accuracy. Robust Sparse CCA performs best and clearly outperforms Robust

Table 6.2: *Simulation results. Average of the angles between the space spanned by the true and estimated canonical vectors; average true positive rate and true negative rate are reported for each method.*

Design	Method	No contamination			t-distribution			Contamination		
		$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR
Sparse	CCA	0.11	1.00	0.00	0.22	1.00	0.00	0.38	1.00	0.00
Low-	Robust CCA	0.14	1.00	0.00	0.15	1.00	0.00	0.15	1.00	0.00
Dimensional	Sparse CCA	0.04	0.98	0.97	0.19	0.94	0.63	0.34	1.00	0.04
	Robust Sparse CCA	0.04	1.00	0.82	0.11	1.00	0.52	0.05	1.00	0.76
NonSparse	CCA	0.08	1.00	NA	0.32	1.00	NA	0.20	1.00	NA
Low-	Robust CCA	0.11	1.00	NA	0.12	1.00	NA	0.12	1.00	NA
Dimensional	Sparse CCA	0.41	0.93	NA	0.67	0.82	NA	0.23	1.00	NA
	Robust Sparse CCA	0.16	0.99	NA	0.22	0.99	NA	0.13	1.00	NA
Sparse										
High-	Sparse CCA	0.65	0.62	0.99	0.70	0.71	0.87	0.36	1.00	0.80
Dimensional	Robust Sparse CCA	0.66	0.84	0.79	0.56	0.82	0.86	0.16	0.96	0.97
NonSparse										
High-	Sparse CCA	0.20	0.74	NA	0.34	0.83	NA	0.33	0.62	NA
Dimensional	Robust Sparse CCA	0.37	0.59	NA	0.34	0.68	NA	0.10	0.99	NA

CCA: for instance, Robust Sparse CCA achieves average estimation accuracy of 0.05 against 0.15 for the contamination setting, see Table 6.2. The non-robust estimators CCA and Sparse CCA are clearly influenced by the outliers, as reflected by the much higher values of the average angle $\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$ in Table 6.2. Sparse CCA now performs even worse than Robust CCA. The considered contamination induces overfitting in Sparse CCA, reflected in the low values of the true negative rate.

Similar conclusions can be drawn in the NonSparse Low-dimensional design. Note that the true negative rate in Table 6.2 is omitted since the true canonical vectors are non-sparse. In the situation without contamination, the price the sparse methods pay is a decreased estimation accuracy, as measured by the average angle. For Robust Sparse CCA compared to Robust CCA this decrease is marginal. In the contaminated settings, the robust methods perform best and show similar performance.

In the High-dimensional designs, only Sparse CCA and Robust Sparse CCA can be performed. In the scenarios without contamination, Sparse CCA performs best. Sparse CCA is, however, closely followed by Robust Sparse CCA both in terms of average estimation accuracy and sparsity recognition performance in the Sparse High-dimensional design. When adding contamination, the performance of Sparse CCA gets distorted. For the heavier tailed t-distribution, the average

estimation accuracy of Robust Sparse CCA compared to Sparse CCA is much better in the Sparse High-dimensional design: 0.56 against 0.70, and comparable in the NonSparse High-dimensional design: 0.34 for both. For the contamination setting, the average estimation accuracy of Robust Sparse CCA is even more than twice as good as the average estimation accuracy of Sparse CCA in both the Sparse and NonSparse High-dimensional design.

In sum, Robust Sparse CCA shows the best overall performance in this simulation study. It performs best in sparse settings where contamination is present. In sparse non-contaminated settings, Robust Sparse CCA is competitive to Sparse CCA. In contaminated non-sparse settings, Robust Sparse CCA is competitive to Robust CCA.

6.5 Applications

We consider three biometric applications. The first data set is low-dimensional and often used in Robust Statistics. The other two data sets are high-dimensional and have been used before in papers on sparse CCA. We show that the performance of Robust Sparse CCA on these data sets is much better than the performance of Sparse CCA.

We compare the performance of the different CCA methods. To decide on the number of canonical variate pairs to extract, we use the maximum eigenvalue ratio criterion, as discussed in Section 6.3. To compare the performance of the CCA approaches, we perform a leave-one-out cross-validation exercise and compute the cross-validation score

$$CV = \frac{1}{r}\frac{1}{h}\sum_{i=1}^{h} ||\widehat{\mathbf{A}}_{-i}^{T}\mathbf{x}_i - \widehat{\mathbf{B}}_{-i}^{T}\mathbf{y}_i||^2, \tag{6.6}$$

where $\widehat{\mathbf{A}}_{-i}^{T}$ and $\widehat{\mathbf{B}}_{-i}^{T}$ contain the estimated canonical vectors when the i^{th} observation is left out of the estimation sample and $h = \lfloor n(1-\alpha)\rfloor$, with $\alpha = 0$ (0% Trimming) or $\alpha = 0.1$ (10% Trimming). We use trimming to eliminate the effect of outliers in the cross-validation score. The method that achieves the lowest cross-validation score has the best out-of-sample performance.

6.5.1 Evaporation data set

We analyze an environmental data set from Freund [1979]. Ten variables (maximum, minimum and average soil temperature; maximum, minimum and average

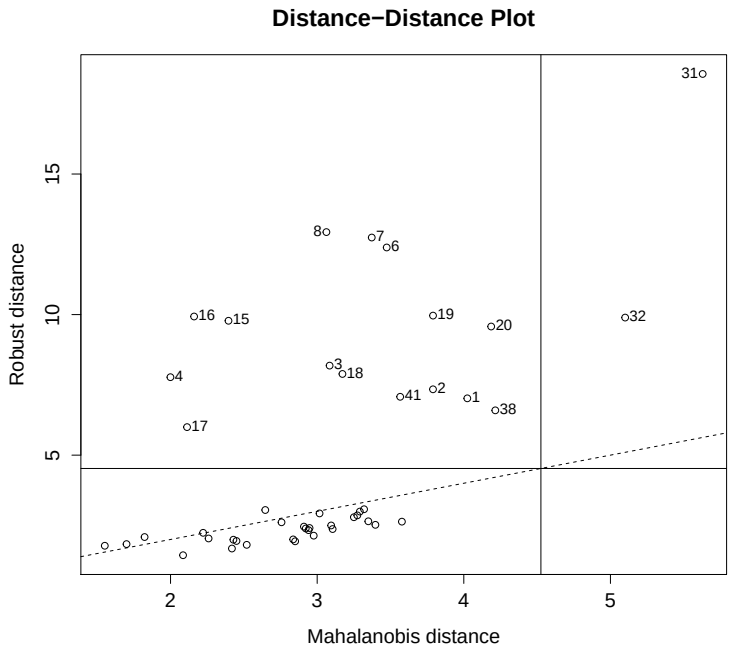


Figure 6.1: *Evaporation data set: Distance-Distance Plot.*

air temperature; maximum, minimum and average daily relative humidity; and total wind) have been measured on $n = 46$ consecutive days from June 6 until July 21. The aim is to find and quantify the relations between the soil temperature variables and the remaining variables.

As a first inspection of the data, we use the Distance-Distance plot [Rousseeuw and van Zomeren, 1990] in Figure 6.1. The Distance-Distance plot displays the robust distances versus the Mahalanobis distances. The vertical and horizontal lines are drawn at values equal to the square root of the 97.5% quantile of a chi-squared distribution with 10 degrees of freedom. Points beyond those lines would be considered as outliers. The Distance-Distance plot reveals some outliers: objects 31 and 32, for example, are extreme outliers. This suggests the need for a robust CCA method. Table 6.3 reports the cross-validation scores from equation (6.6) for the four CCA methods. For all methods two canonical variate pairs are extracted. Robust Sparse CCA achieves the best cross-validation score.

Table 6.4 shows the estimated canonical vectors for the Robust CCA and

Table 6.3: *Evaporation data set: Cross-validation score for standard CCA, Robust CCA, Sparse CCA and Robust Sparse CCA.*

Method	CV-score 0% Trimming	CV-score 10% Trimming
CCA	0.74	0.49
Robust CCA	0.57	0.39
Sparse CCA	0.57	0.41
Robust Sparse CCA	0.48	0.31

Table 6.4: *Evaporation data set: Estimated canonical vectors using Robust CCA and Robust Sparse CCA.*

		Robust CCA		Robust Sparse CCA	
	Variables \ Canonical Vectors	1	2	1	2
First	MAXST: Max. daily soil temperature	-0.35	-0.76	0	-0.70
data	MINST: Min. daily soil temperature	0.03	0.63	0	0.71
set	AVST: Avg. daily soil temperature	0.93	0.18	1	0
Second	MAXAT: Max. daily air temperature	0.54	-0.11	0.94	0
data	MINAT: Min. daily air temperature	0.67	0.84	0.14	0.38
set	AVAT: Avg. daily air temperature	0.14	-0.03	0.17	0.36
	MAXH: Max. daily relative humidity	-0.13	0.09	0	0
	MINH: Min. daily relative humidity	-0.03	0.36	0	0.85
	AVH: Avg. daily relative humidity	-0.28	0.32	-0.24	0
	WIND: Total wind, measured in miles per day	-0.37	-0.19	0	0
	Canonical correlations	0.93	0.56	0.87	0.48

Robust Sparse CCA method. By adding the penalty term, the number of non-zero coefficients is reduced from 20 to 10. The price to pay for the sparseness is a slight decrease in the estimated canonical correlations (computed using the bivariate MCD estimator, see Section 6.3): they drop from 0.93 to 0.87 for the first one, and from 0.56 to 0.48 for the second canonical correlation. We find this decrease acceptable, given the gained sparsity in the canonical vectors. The sparse structure of the canonical vectors facilitates interpretation. The first canonical variate in the soil temperature data set, for instance, is uniquely determined by the variable AVST.

Table 6.5: *Nutrimouse data set: Cross-validation score for Sparse CCA and Robust Sparse CCA.*

Method	CV-score 0% Trimming	CV-score 10% Trimming
Sparse CCA	98.78	92.53
Robust Sparse CCA	6.30	4.31

6.5.2 Nutrimouse data set

This genetic data set is publicly available in the R package `CCA` [Gonzalez et al., 2008]. Two sets of variables, i.e. gene expressions and fatty acids, are available for $n = 40$ mice. The first set contains expressions of $p = 120$ genes measured in liver cells. The second set of variables contains concentrations of $q = 21$ hepatic fatty acids (FA). In this experiment, there are two groups of mice (wild-type and PPARα deficient mice) that receive a specific diet (five possible diets). More details on how the data were obtained can be found in Martin et al. [2007]. The aim is to identify a small set of genes which are correlated with the fatty acids.

In this data set, the number of experimental units is smaller than the number of variables. Therefore, standard CCA nor robust CCA can be performed. Robust Sparse CCA and Sparse CCA can be applied in this high-dimensional setting and produce interpretable, sparse canonical vectors. For both methods, one canonical variate pair is extracted. The cross-validation scores from equation (6.6) are reported in Table 6.5. Robust Sparse CCA outperforms Sparse CCA. The cross-validation scores are reduced by about 90% when using the robust method.

Next, we discuss the estimated canonical vectors obtained using the Robust Sparse CCA method. The top panel of Figure 6.2 displays the coefficients of the selected genes, i.e. those genes with non-zero estimated coefficients, in the first canonical vector: 24 out of 120 variables are selected. The solution is very sparse, facilitating interpretation. Martin et al. [2007] find a consistent reduction of `Cyp3a11` in PPARα livers on the one hand, and an overexpression of `CAR1` on the other hand. Both genes are selected and have among the highest (absolute) coefficients. The coefficients of the selected fatty acids are displayed in the bottom panel of Figure 6.2: 13 out of 21 fatty acid variables are selected. The fatty acids `C22:6n-3`, `C22:5n-3`, `C22:5n-6`, `C22:4n-3` and `C20:5n-3` are related to the effect of the five diets used in this experiment. From Figure 6.2, we see that four out of these five fatty acids are selected.

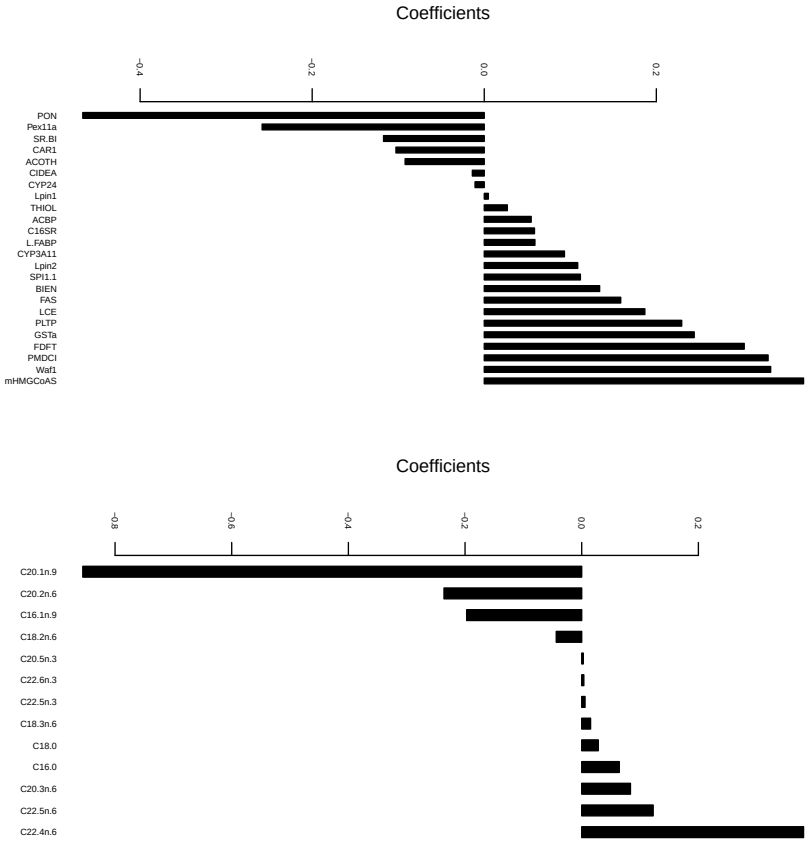


Figure 6.2: *Nutrimouse data set: Coefficients of selected genes (top) and coefficients of selected fatty acids (bottom) in the first canonical vector pair.*

6.5.3 Breast cancer data set

The genetic data set is described in Chin et al. [2006] and available in the `R package PMA` [Witten et al., 2011]. Two sets of data, i.e. gene expression data (19 672 variables) and comparative genomic hybridization (CGH) data (2149 variables) are available for $n = 89$ patients, and this for 23 chromosomes. We analyze the data for each of the chromosomes separately, each time using the CGH and gene expression variables for that particular chromosome. Depending on the chromosome, either 1, 2, 3, or 4 canonical vector pairs are extracted. The aim is to identify a subset of CGH variables that are correlated with a subset of gene expression variables.

Results of the cross-validation scores of equation (6.6) are reported in Figure 6.3. For each of the 23 chromosomes, we plot the value of the cross-validation score (0% trimming) for Robust Sparse CCA (horizontal axis) and Sparse CCA (vertical axis). Results when using 10% trimming are similar and, therefore, omitted. The cross-validation scores of Robust Sparse CCA are much better than those of Sparse CCA: all points are lying above the 45°-line. For chromosomes 1, 3, 4, and 11, for instance, the cross-validation scores of Robust Sparse CCA are more than 10 times lower than those of Sparse CCA. Since Robust Sparse CCA performs much better, outliers might be present for these chromosomes. Hence, it is safer to use Robust Sparse CCA instead of Sparse CCA.

We use the estimates from Robust Sparse CCA to detect the outliers. To this end, we create the Residual Distance plot of the residuals $\mathbf{X}\widehat{\mathbf{A}} - \mathbf{Y}\widehat{\mathbf{B}}$, and this for each of the 23 chromosomes. The Residual Distance plot displays the robust distance of the residuals (vertical axis) versus the observation number (horizontal axis). Points above the horizontal black line are marked as outliers. Results for chromosome 3 and 8 are displayed in Figure 6.4, results for the other chromosomes are available upon request. For some chromosomes, like chromosome 3, the difference in cross-validation scores of Robust Sparse CCA and Sparse CCA in Figure 6.3 is outspoken, suggesting that outliers might be present. We use the Residual Distance plot (Figure 6.4, left panel) to detect which patients are outlying. In the Residual Distance plot of chromosome 3 a lot of patients are marked as outliers. For chromosome 8, on the other hand, the cross-validation scores of Sparse CCA and Robust Sparse CCA are nearly identical, which might suggest that there are no outliers. Looking at the Residual Distance Plot of chromosome 8 (Figure 6.4, right panel), no outliers are indeed detected.

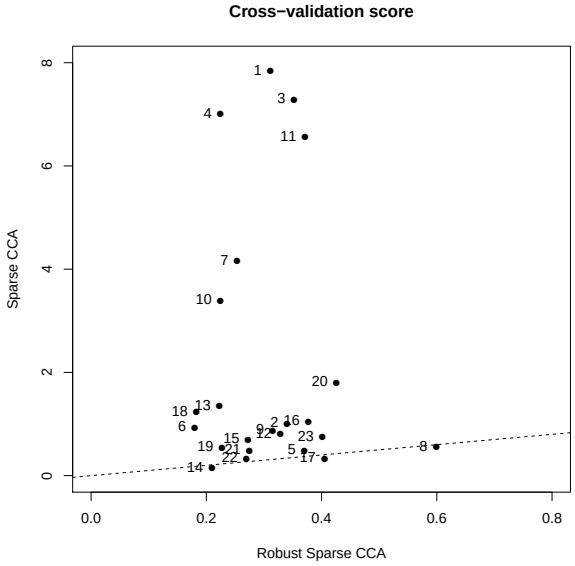


Figure 6.3: *Breast cancer data set: 23 cross-validation scores (one for each chromosome) for Robust Sparse CCA (horizontal axis) and Sparse CCA (vertical axis). The dashed line is the 45°-line.*

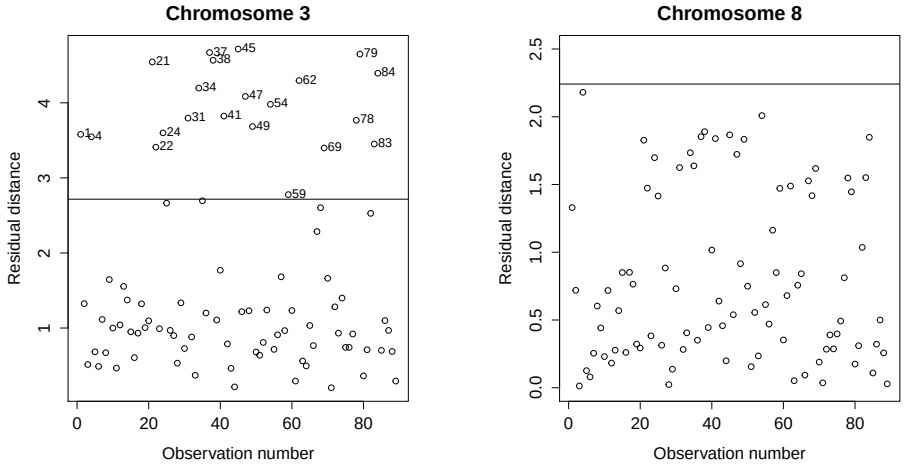


Figure 6.4: *Breast cancer data set: Residual Distance Plot for chromosome 3 (left) and chromosome 8 (right).*

6.6 Discussion

Sparse Canonical Correlation Analysis delivers interpretable canonical vectors, with some of its elements estimated as exactly zero. Robust Sparse CCA retains this advantage, while at the same coping with outlying observations.

The canonical vectors are given by the eigenvectors of two particular matrices, see for instance Johnson and Wichern (1998, Chapter 10). Typically, the canonical vectors are estimated by taking the sample versions of those covariance matrices and computing the corresponding eigenvectors. One could think of estimating those covariance matrices with an estimator that is robust and sparse at the same time, and then, to compute the eigenvectors. This approach, however, would results in canonical vectors being non-sparse. To circumvent this pitfall, we reformulate the CCA problem in a regression framework. A simulation study and three biometric examples show the advantages of the Robust Sparse CCA method over its benchmarks.

Robust Sparse CCA has three important advantages over Robust CCA. (i) Robust Sparse CCA improves model interpretation since only a limited number of variables, those corresponding to the non-zero elements of the canonical vectors, enter the estimated canonical variates (cfr. evaporation application), (ii) if the number of variables approaches the sample size, the estimation precision of Robust CCA suffers, and (iii) if the number of variables exceeds the sample size, Robust CCA can not even be performed. Robust Sparse CCA can still be computed (cfr. nutrimouse and breast cancer application).

Several questions are left for future research. One could use a joint selection criterion for the number of canonical variate pairs and the sparsity parameter. This would, however, increase computation time substantially. To induce sparsity in the canonical vectors, we use a Lasso penalty. Other penalty functions such as the Adaptive Lasso [Zou, 2006] could be considered. The Adaptive Lasso is consistent for variable selection, whereas the Lasso is not. Furthermore, we use a regularized version of the LTS estimator. One could also use a regularized version of the S-estimator or the MM-estimator to increase efficiency. Up to our knowledge, however, the sparse LTS is the only robust sparse regression estimator for which efficient code [Alfons, 2014] is available.

Outlook

This thesis discusses several aspects of sparse estimation for high-dimensional time series models. Such sparse high-dimensional model estimation is an active area of research see, for instance, the recent work on graphical VAR models [Wild et al., 2010], Bayesian graphical VAR models [Ahelegbey et al., 2016], or theoretical properties of sparse estimators for high-dimensional time series models [Basu and Michaelidis, 2015]. This constant flow of developments opens up new possibilities for future research.

Several questions on sparse model estimation remain open. Regarding the selection of the sparsity parameter, information-based criteria such as BIC should be compared to cross-validation procedures or stability selection procedures [Meinshausen and Bühlmann, 2010]. For the bootstrap procedure from Chapter 3, its performance when the sparsity parameter is re-selected in each bootstrap run (as is currently the case) should be compared to its performance when the sparsity parameter is kept fixed. Besides, the time series models from Chapters 1 to 4 assume a common lag structure for all included time series. The use of hierarchical penalties [Bien et al., 2013], that allow each time series to have its own lag structure, should be investigated. Also the dependency of the penalized maximum likelihood procedures on the normality of the error terms could be relaxed in future research. A large comparative study on the forecast performance of sparse estimators compared to other commonly used forecast techniques such as Partial Least Squares or Boosting could further deepen our understanding on the forecast abilities of sparse estimators.

Furthermore, it would be interesting to consider sparse *multi-class* estimation. The aim would be to jointly estimate K models corresponding to K distinct but related classes. In Chapter 1, we estimate 15 separate VAR models, one for each store. Since the stores belong to the same retailer, one might expect the K models to be similar to each other. Therefore, the multi-class estimator would encourage

(i) many elements of the autoregressive matrices to be identical across classes, and (ii) shared sparsity patterns across classes. At the same time, there might exist important differences between the stores stemming from the specific shopping behavior in each store. Hence, the multi-class estimator should allow for small differences between classes. It would be interesting to investigate the influence on the results when jointly estimating the K high-dimensional VAR models.

Sparse multi-class estimation would also be highly relevant to the CCA framework from Chapters 5 and 6. It allows for a direct comparison between patients (class one) and controls (class two). The multi-class estimation could result in a more precise estimation of the associations and a deeper understanding of the differences in genetic associations between patients and controls.

The novelty of combining robust and sparse estimation, as in Chapter 6, also calls for future research. If the outliers detected by the robust estimator form a group on their own, mixture models are worth considering. Besides, while several *rowwise* robust sparse estimators have already been developed, less work has been done in sparsifying the recently introduced *cellwise* robust estimators. Rowwise robust estimators flag either a whole *row* (observation) in the data set as outlying or not. This leads, however, to a considerable loss of information in high-dimensional data sets. Cellwise robust estimators, in contrast, only flag a *cell* of the data set as outlying or not. As such, sparse cellwise robust estimators could provide a valuable alternative to sparse rowwise robust estimators in high-dimensions.

List of Figures

List of Tables

Bibliography

D. Aaker and K. Keller. Consumer evaluations of brand extensions. *Journal of Marketing*, 54(1):27–41, 1990.

K. Abberger. Qualitative business surveys and the assessment of employment – A case study for Germany. *International Journal of Forecasting*, 23(2):377–389, 2007.

F. Abegaz and E. Wit. Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, 14(3):586–599, 2013.

J.G. Adrover and S.M. Donato. A robust predictive approach for canonical correlation analysis. *Journal of Multivariate Analysis*, 133:356–376, 2015.

D. F. Ahelegbey, M. Billio, and R. Casarin. Bayesian graphical models for structural vector autoregressive processes. *Journal of Applied Econometrics*, 31(2): 357–386, 2016.

K.L. Ailawadi, J.P. Beauchamp, N. Donthu, D.K. Gauri, and V. Shankar. Communication and promotion decisions in retailing: A review and directions for future research. *Journal of Retailing*, 85(1):42–55, 2009.

A. Ainslie and P.E. Rossi. Similarities in choice behavior across product categories. *Marketing Science*, 17(2):91–106, 1998.

P. Albuquerque, B.J. Bronnenberg, and C. Corbett. A spatio-temporal analysis of global diffusion of iso 9000 and iso 14000 certification. *Management Science*, 53(3):451–468, 2007.

A. Alfons. *robustHD: Robust methods for high-dimensional data*, 2014. R package version 0.5.0.

A. Alfons, C. Croux, and S. Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7:226–248, 2013.

A. Alfons, C. Croux, and S. Gelper. Robust groupwise least angle regression. *Computational Statistics and Data Analysis*, 93:421–435, 2016.

A. Alonso, H. Geys, G. Molenberghs, M.G. Kenward, and T. Vangeneugden. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal*, 45(8):931–945, 2003.

B.G. An, J. Guo, and H. Wang. Multivariate regression shrinkage and selection by canonical correlation analysis. *Computational Statistics and Data Analysis*, 62:93–107, 2013.

T.W. Anderson. *An introduction to multivariate statistical analysis.* John Wiley & Sons, Inc., New York, 1958.

S. Angilella and S. Mazzu. The financing of innovative smes: A multicriteria credit rating model. *European Journal of Operational Research*, 244(2):540–554, 2015.

S. Asur and B.A. Huberman. Predicting the future with social media. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 1:492–499, 2010.

H. Baghestani. Cointegration analysis of the advertising-sales relationship. *Journal of Industrial Economics*, 39(6):671–681, 1991.

M.T. Bahadori and Y. Liu. An examination of practical granger causality inference. *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013.

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317, 2008.

M. Ballings and D. Van den Poel. CRM in social media: Predicting increases in Facebook usage frequency. *European Journal of Operational Research*, 244(1): 248–260, 2015.

M. Banbura, D. Giannone, and L. Reichlin. Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

S. Bandyopadhyay. A dynamic model of cross-category competition: Theory, tests and applications. *Journal of Retailing*, 85(4):468–479, 2009.

M. Barogozzi, M. Lippi, and M. Luciani. Non-stationary dynamic factor models for large datasets. *arXiv*, 1602.0239v1, 2016.

S. Basu and G. Michaelidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

S. Basuroy, M.K. Mantrala, and R.G. Walters. The impact of category management on retailer prices and performance: Theory and evidence. *Journal of Marketing*, 65(4):16–32, 2001.

T. Beck and A. Demirguc-Kunt. Small and medium-size enterprises: Access to finance as a growth constraint. *Journal of Banking & Finance*, 30(11):2931–2943, 2006.

D.R. Bell, T.H. Ho, and C.S. Tang. Determining where to shop: Fixed and variable costs of shopping. *Journal of Marketing Research*, 35(3):352–369, 1998.

E. Bernardini and G. Cubadda. Macroeconomic forecasting and structural analysis through regularized reduced-rank regression. *International Journal of Forecasting*, 31(3):682–691, 2015.

R. Bezawada, S. Balachander, P.K. Kannan, and V. Shankar. Cross-category effects of aisle and display placements: a spatial modeling approach and insights. *Journal of Marketing*, 73(3):99–117, 2009.

J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013.

N. Blasco, P. Corredor, C. Del Rio, and R. Santamaria. Bad news and Dow Jones make the Spanish stocks go round. *European Journal of Operational Research*, 163(1):253–275, 2005.

R.C. Blattberg, E.J. Fox, and M.E. Purk. *Category management: A series of implementation Guides. Vols. I-IV.* Food Marketing Institute, Washington ,DC, 1995.

A. Bonfrer, E.R. Berndt, and A. Silk. Anomalies in estimates of cross-price elasticities for marketing mix models: Theory and empirical test. *National Bureau of Economic Research*, NBER Working Paper 12756, 2006.

J. Branco, C. Croux, P. Filzmoser, and M. Oliveira. Robust canonical correlations: A comparative study. *Computational Statistics*, 20(2):203–229, 2005.

J. Breitung and G. Cubadda. Testing for cointegration in high-dimensional systems. *CEIS Working Paper*, 2011.

R.A. Briesch, W.R. Dillon, and E.J. Fox. Category positioning and store choice: The role of destination categories. *Marketing Science*, 32(3):488–509, 2013.

D.R. Brillinger. *Time Series: Data analysis and theory.* Holt, Rinehart, and Winston, New York, 1975.

G. Bruno. Consumer confidence and consumption forecast: a non-parametric approach. *Empirica*, 41(1):37–52, 2014.

P. Bühlmann and T. Hothorn. Twin boosting: Improved feature selection and prediction. *Statistics and Computing*, 20(2):119–138, 2010.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data. Methods, Theory and Applications.* Springer, 2011.

F. Bunea, Y. She, and M.H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

A. Carriero, G. Kapetanios, and M. Marcellino. Forecasting large datasets with Bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, 26(5):735–761, 2011.

A. Carriero, G. Kapetanios, and M. Marcellino. Forecasting government bond yields with large bayesian vector autoregressions. *Journal of Banking & Finance*, 36(7):2026–2047, 2012.

G. Cavaliere, A. Rahbek, and A.M. R. Taylor. Bootstrap determination of the co-integration rank in vector autoregressive models. *Econometrica*, 80(4):1721–1740, 2012.

A. Chatterjee and S.N. Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

L. Chen and J.Z. Huang. Sparse reduced-rank regression with covariance estimation. *Statistical Computing*, 26:461–470, 2016.

Z-Y. Chen, Z-P. Fan, and M. Sun. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research*, 223:462–472, 2012.

K. Chin, S. DeVries, J. Fridlyand, P. Spellman, R. Roydasgupta, W.L. Kuo, A. Lapuk, R. Neve, Z. Qian, T. Ryder, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541, 2006.

P. Chintagunta, D.M. Hanssens, and J. Hauser. Call for papers. marketing science special issue on big data: Integrating marketing, statistics, and computer science. *Marketing Science*, 32(4):678, 2013.

C. Christiansen, J.N. Eriksen, and S.V. Moller. Forecasting US recessions: The role of sentiment. *Journal of Banking & Finance*, 49:459–468, 2014.

O. Claveria, E. Pons, and R. Ramos. Business and consumer expectations and macroeconomic forecasts. *International Journal of Forecasting*, 23(1):47–69, 2007.

J.A. Cotsomitis and C.C. Kwan. Can consumer confidence forecast household spending? Evidence from the European Commission busniness and consumer surveys. *Southern Economic Journal*, 72:597–610, 2006.

R. Cruz-Cano and M.-L.T. Lee. Fast regularized canonical correlation analysis. *Computational Statistics and Data Analysis*, 70:88–100, 2014.

R. Davidson and J.G. McKinnon. Graphical methods for investigating the size and power of hypothesis tests. *The Manchester School*, 66:1–26, 1998.

R.A. Davis, P. Zang, and T. Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, DOI:10.1080/10618600.2015.1092978, 2015.

A. Deaton and J. Muellbauer. *Economics and consumer behavior*. Cambridge University Press, New York, 1980.

C. Dehon and C. Croux. Analyse canonique basée sur des estimateurs robustes de la matrice de covariance. *La Revue de Statistique Appliquée*, 2:5–26, 2002.

M.G. Dekimpe and D.M. Hanssens. The persistence of marketing effects on sales. *Marketing Science*, 14(1):1–21, 1995.

M.G. Dekimpe and D.M. Hanssens. Sustained spending and persistent response: A new look at long-term marketing profitability. *Journal of Marketing Research*, 36(4):397–412, 1999.

M.G. Dekimpe and D.M. Hanssens. Time-series models in marketing: Past, present and future. *International Journal of Research in Marketing*, 17:183–193, 2000.

M.G. Dekimpe, D.M. Hanssens, and J.M. Silva-Risso. Long-run effects of price promotions in scanner markets. *Journal of Econometrics*, 89(1-2):269–291, 1999.

G. Dell'Ariccia, E. Detragiache, and R. Rajan. The real effects of banking crises. *Journal of Financial Intermediation*, 17:89–112, 2008.

S.K. Dhar, S.J. Hoch, and N. Kumar. Effective category management depends on the role of the category. *Journal of Retaling*, 77(2):165–184, 2001.

F.X. Diebold and R.S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:253–263, 1995.

D.L. Donoho and J.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

C. Dreger and D.A. Kholodilin. Forecasting private consumption by consumer surveys. *Journal of Forecasting*, 32(1):10–18, 2013.

R.Y. Du and A. Kamakura. Where did all that money go? understanding how consumers allocate their consumption budget. *Journal of Marketing*, 72(6): 109–131, 2008.

T. Elrod, G.J. Russell, A.D. Shocker, R.L. Andrews, L. Bacon, B.L. Bayus, J.D. Carroll, R.M. Johnson, W.A. Kamakura, P. Lenk, J.A. Mazanec, V.R. Rao, and V. Shankar. Inferring market structure from customer response to competing and complementary products. *Marketing Letters*, 13(3):221–232, 2002.

R.F. Engle and C.W.J. Granger. Cointegration and error correction - representation, estimation, and testing. *Econometrica*, 55:251–276, 1987.

T. Engsted and C. Tanggaard. Cointegration and the US term structure. *Journal of Banking and Finance*, 18:167–181, 1994.

T. Erdem. An empirical analysis of umbrella branding. *Journal of Marketing Research*, 35(3):339–351, 1998.

T. Erdem and B. Sun. An empirical investigation of the spillover effects of advertising and sales promotions in umbrella branding. *Journal of Marketing Research*, 39(4):408–420, 2002.

P. Fader and L.M. Lodish. A cross-category analysis of category structure and promotional activity for grocery products. *Journal of Marketing*, 54(4):52–65, 1990.

J. Fan, J. Lv, and L. Qi. Sparse high-dimensional models in economics. *Annual Review of Economics*, 3:291–317, 2011.

J.Q. Fan and R.Z. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001.

A.I. Fernandez, F. Gonzalez, and N. Suarez. The real effect of banking crises: Finance or asset allocation effects? Some international evidence. *Journal of Banking & Finance*, 37(7):2419–2433, 2013.

C. Fornell, R.T. Rust, and M.G. Dekimpe. The effect of customer satisfaction on consumer spending growth. *Journal of Marketing Research*, 47(1):28–35, 2010.

T. Foucart. Multiple linear regression on canonical correlation variables. *Biometrical Journal*, 41(5):559–572, 1999.

R.J. Freund. Multicollinearity etc. some 'new' examples. *American Statistical Association Proceedings of Statistical Computing Section*, pages 111–112, 1979.

J. Friedman. Fast sparse regression and classification. *International Journal of Forecasting*, 28(3):722–738, 2012.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. *glasso: Graphical lasso- estimation of Gaussian graphical models*, 2011. URL `http://www-stat.stanford.edu/ tibs/glasso`. R package version 1.4.

A. Fujita, J.R. Sato, H.M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M.C. Sogayar, and C.E. Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1:No. 39, 2007.

M. Gangwar, N. Kumar, and R.C. Rao. Consumer stockpiling and competitive promotional strategies. *Marketing Science*, 33(1):94–113, 2014.

D. Gefang. Bayesian double adaptive elastic-net lasso for var shrinkage. *International Journal of Forecasting*, 30(1):1–11, 2014.

S. Gelper and C. Croux. On the construction of the european economic sentiment indicator. *Oxford Bulletin of Economics and Statistics*, 72(1):47–62, 2010.

S. Gelper, I. Wilms, and C. Croux. Identifying demand effects in a large network of product categories. *Journal of Retailing*, 92(1):25–39, 2016.

S. Ghosh, M.D. Troutt, J.H. Thornton, and O.F. Offodile. An empirical method for assessing the research relevance gap. *European Journal of Operational Research*, 201:942–948, 2010.

J.V. Giese. Level, slope, curvature: Characterising the yield curve in a cointegrated var model. *Economics*, 2. No. 2008-28, 2008.

I. Gonzalez and S. Dejean. *Canonical correlation analysis*, 2009. R package version 1.2.

J. Gonzalez, C. Sismeiro, S. Dutta, and P. Stern. Can branded drugs benefit from generic entry? the role of detailing and price in switching to non-bioequivalent moleculus. *International Journal of Research in Marketing*, 25:247–260, 2008.

J. Graffelman and F. van Eeuwijk. Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research. *Biometrical Journal*, 47(6):863–879, 2005.

S. Gross, B. Narasimhan, R. Tibshirani, and D.M. Witten. *Correlate: Sparse canonical correlation analysis for the integrative analysis of genomic data*, 2011. URL `http://www-stat.stanford.edu/ tibs/Correlate/correlate.pdfl`. User guide and technical document.

J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1991.

J. Hansson, P. Jansson, and M. Lof. Business survey data: Do they help in forecasting GDP growth? *International Journal of Forecasting*, 21:377–389, 2005.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2nd edition, 2009. ISBN 978-0-387-84857-0.

A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

G. Hommel and S. Kropf. Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biometrical Journal*, 47(4):554–562, 2005.

C. Horvath and D. Fok. Moderating factors of immediate, gross, and net cross-brand effects of price promotions. *Marketing Science*, 32(1):127–152, 2013.

C. Horvath, P.S.H. Leeflang, J.E. Wieringa, and D.R. Wittink. Competitive reaction- and feedback effects based on varx models of pooled store data. *International Journal of Research in Marketing*, 22(4):415–426, 2005.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

N.-J. Hsu, H.-L. Hung, and Y.-M.. Chang. Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis*, 52(7): 3645–3657, 2008.

T. Huang, R. Fildes, and D. Soopramanien. The value of competitive information in forecasting fmcg retail product sales and the variable selection problem. *European Journal of Operational Research*, 237:738–748, 2014.

R.J. Hyndman. *forecast: Forecasting functions for time series and linear models*, 2014. URL `http://cran.r-project.org/web/packages/forecast/forecast.pdf`. R package version 5.2.

R. Iaci, T. N. Sriram, and X. Yin. Multivariate association and dimension reduction: A generalization of canonical correlation analysis. *Biometrics*, 66(4): 1107–1118, 2010.

A. Inoue and L. Kilian. How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association*, 103(482):511–522, 2008.

A.J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

S. Johansen. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2-3):231–254, 1988.

S. Johansen. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59:1551–1580, 1991.

S. Johansen. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press, Oxford, 1996.

S. Johansen. A small sample correction of the test for cointegration rank in the vector autoregressive model. *Econometrica*, 70(5):1929–1961, 2002.

S. Johansen, R. Mosconi, and B. Nielsen. Cointegration analysis in the presence of structural breaks in the deterministic trend. *Econometrics Journal*, 3:216–249, 2000.

R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, London, 1998.

W.A. Kamkura and W. Kang. Chain-wide and store-level analysis for cross-category management. *Journal of Retailing*, 83(2):159–170, 2007.

L.R. Klein and S. Oezmucur. The use of consumer and business surveys in forecasting. *Economic Modelling*, 27(6):1453–1462, 2010.

G. Koop and D. Korobilis. Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, 3(4):267–358, 2009.

D. Korobilis. VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, 28(2):204–230, 2013.

J.P. Kreiss and S. Lahiri. *Bootstrap methods for time series. In: Rao, T., Rao, S. and Rao, C. (Eds.) Handbook of Statistics 30. Time Series Analysis: Methods and Applications*. North Holland, 2012.

R.S. Kroszner, L. Laeven, and D. Klingebiel. Banking crises, financial dependence and growth. *Journal of Financial Economics*, 84(1):187–228, 2007.

N.L. Kudraszow and R.A. Maronna. Robust canonical correlation analysis: a predictive approach. *Working paper*, 2011.

V. Kumar and T.V. Krishnan. Multinational diffusion models: An alternative approach. *Marketing Science*, 21(3):318–330, 2002.

V. Kumar, R.P. Leone, and J.N. Gaskin. Aggregate and disaggregate sector forecasting using consumer confidence measures. *International Journal of Forecasting*, 11:361–377, 1995.

C. Lam and Q. Yao. Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 40(2):694–726, 2012.

H.O. Lancaster. The combination of probabilities arising from data in discrete distributions. *Biometrika*, 36:370–382, 1949.

J. Lauter, F. Horn, M. Rosolowski, and E. Glimm. High-dimensional data analysis: selection of variables, data compression and graphics - application to gene expression. *Biometrical Journal*, 51(2):235–251, 2009.

S. Lee, J. Kim, and M. Allenby. A direct utility model for asymmetric complements. *Marketing Science*, 32(3):454–470, 2013.

P.S.H. Leeflang and J.P. Selva. Cross-demand effects of price promotions. *Journal of Academy of Marketing Science*, 40(4):572–586, 2012.

P.S.H. Leeflang, J.P. Selva, A. Van Dijk, and D.R. Wittink. Decomposing the sales promotion bump accounting for cross-category effects. *International Journal of Research in Marketing*, 25(3):201–214, 2008.

A. Lemmens, C. Croux, and M.G. Dekimpe. On the predictive content of production surveys: A pan-european study. *International Journal of Forecasting*, 21:363–375, 2005.

P. Lenk and B. Orme. The value of informative priors in bayesian inference with sparse data. *Journal of Marketing Research*, 46(6):832–845, 2009.

A. Levin, C.F. Lin, and C.S.J. Chu. Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, 108(1):1–24, 2002.

Y. Li, B. Nan, and J. Zhu. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2): 354–363, 2015.

Z. Liao and P.C.B. Phillips. Automated estimation of vector error correction models. *Econometric Theory*, 31:581–646, 2015.

R.W. Lissitz, P.H. Schonemann, and J.C. Lingoes. A solution to the weighted procrustes problem in which the transformation is in agreement with the loss function. *Psychometrika*, 41:547–550, 1976.

R.B. Litterman. A bayesian procedure for forecasting with vector autoregression. *Working Paper, Massachusetts Institute of Technology, Dept. of Economics*, 1980.

R.B. Litterman. Forecasting with bayesian vector autoregressions: Five years of experience. *Journal of Business & Economic Statistics*, 4:25–38, 1986.

L.M. Lodish and C.F. Mela. If brands are built over years, why are they managed over quarters? *Harvard Business Review*, 85(10):158, 2007.

H. Lütkepohl. *Introduction to multiple time series analysis*. Springer-Verlag: Berlin-Germany, 1993.

H. Lütkepohl. *New introduction to multiple time series analysis*. Springer-Verlag, 2007.

H. Lütkepohl and M. Kratzig. *Applied time series econometrics*. Cambridge University Press, 2004.

A. Lykou and J. Whittaker. Sparse CCA using a lasso with positivity constraints. *Computational Statistics and Data Analysis*, 54(12):3144–3157, 2010.

Y. Ma, P.B. Seetharaman, and C. Narasimhan. Modeling dependencies in brand choice outcomes across complementary categories. *Journal of Retailing*, 88(1): 47–62, 2012.

P. Manchanda, A. Ansari, and S. Gupta. The shopping basket: A model for multicategory purchase incidence decisions. *Marketing Science*, 18(2):95–114, 1999.

P.G. Martin, H. Guillon, F. Lasserre, S. Dejean, A. Lan, J.M. Pascussi, M. San-Cristobal, P. Legrand, P. Besse, and T. Pineau. Novel aspects of PPARα-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. *Hepatology*, 45(3):767–777, 2007.

K. Martinsen, F. Ravazzolo, and F. Wulfsberg. Forecasting macroeconomic variables using disaggregate survey data. *International Journal of Forecasting*, 30(1):65–77, 2014.

L. Meier. *grplasso: Fitting user specified models with group lasso penalty*, 2009. URL `http://CRAN.R-project.org/package=grplasso`. R package version 0.4-2.

L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72(4):417–473, 2010.

N. Meinshausen, L. Meier, and P. Bühlmann. p-values for high-dimensional regression. *Journal of American Statistical Association*, 104(488):1671–1681, 2009.

S. Musti and R.L. D'Ecclesia. Term structure of interest rates and the expectation hypothesis: The euro area. *European Journal of Operational Research*, 185: 1596–1606, 2008.

B. Nielsen. On the distribution of tests of cointegration. *Econometric Reviews*, 23(1):1–23, 2004.

B. Nielsen and A. Rahbek. Similarity issues in cointegration models. *Oxford Bulletin of Economics and Statistics*, 62(1):5–22, 2000.

V. Nijs, M.G. Dekimpe, J.B. Steenkamp, and D.M. Hanssens. The category demand effects of price promotions. *Marketing Science*, 20(1):1–22, 2001.

V. Nijs, S. Srinivasan, and K. Pauwels. Retail-price drivers and retailer profits. *Marketing Science*, 26(4):473–487, 2007.

R. Niraj, V. Padmanabhan, and P.B. Seetharaman. A cross-category model of households' incidence and quantity decisions. *Marketing Science*, 27(2):225–235, 2008.

G. Oestreicher-Singer, B. Libai, L. Sivan, E. Carmi, and O. Yassin. The network value of products. *Journal of Marketing*, 77(3):1–14, 2013.

R. Östermark. Multivariate cointegration analysis ofthe finnish-japanese stock markets. *European Journal of Operational Research*, 134(3):498–507, 2001.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8:1–34, 2009.

B.P. Pashigian. *Price theory and applications*. McGraw-Hill Higher Education, 1998.

K. Pauwels. How retailer and competitor decisions drive the long-term effectiveness of manufacturer promotions for fast moving consumer goods. *Journal of Retailing*, 83(3):297–308, 2007.

K. Pauwels, D.M. Hanssens, and S. Siddarth. The long-term effects of price promotions on category incidence, brand choice and purchase quantity. *Journal of Marketing Research*, 39(4):421–439, 2002.

J. Peng, J. Zhu, A. Bergamaschi, W. Han, D. Y. Noh, J. R. Pollack, and P. Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1):53–77, 2010.

H.H. Pesaran and S. Shin. Generalized impulse response analysis in linear multivariate models. *Economics Letters*, 58(1):17–29, 1998.

P.C.B. Phillips and S. Ouliaris. Asymptotic properties of residual based tests for cointegration. *Econometrica*, 58(1):165–193, 1990.

D.N. Politis and J.P. Romano. The stationary bootstrap. *Journal of Americal Statistical Association*, 89(428):1303–1313, 1994.

C. Prabhakar and B.L. Fridley. Comparison of penalty functions for sparse canonical correlation analysis. *Computational Statistics and Data Analysis*, 56(2): 245–254, 2012.

S. Pradhan. *Retailing management: Text & Cases*. Tata McGraw-Hill Education, New Delhi, 2009.

S. Rajagopalan and N. Xia. Product variety, pricing and differentiation in a supply chain. *European Journal of Operational Research*, 217(1):84–93, 2012.

A. Rothman, E. Levina, and J. Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4): 947–962, 2010.

P. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

P. Rousseeuw and B.C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990.

G.J. Russell and W.A. Kamakura. Modeling multiple category brand preference with household basket data. *Journal of Retailing*, 73(4):439–461, 1997.

G.J. Russell and A. Petersen. Analysis of cross category dependence in market basket selection. *Journal of Retailing*, 76(3):367–392, 2000.

G.J. Russell, S. Ratneshwar, A.D. Shocker, D. Bell, A. Bodapati, A. Degeratu, L. Hildebrandt, N. Kim, S. Ramaswami, and V.H. Shankar. Multiple-category decision making: review and synthesis. *Marketing Letters*, 10(3):319–332, 1999.

B.K. Sahoo and D. Acharya. An alternative approach to monetary aggregation in dea. *European Journal of Operational Research*, 204(3):672–682, 2010.

H. Schwender, K. Ickstadt, and J. Rahnenfuehrer. Classification with high-dimensional genetic data: Assigning patients and genetic features to known classes. *Biometrical Journal*, 50(6):911–926, 2008.

E. Shafir, P. Diamond, and A. Tversky. Money illusion. *Quarterly Journal of Economics*, 112(2):341–374, 1997.

V. Shankar and P.K. Kannan. An across-store analysis of intrinsic and extrinsic cross-category effects. *Customer Needs and Solutions*, 1:143–153, 2014.

S.J. Shin and Y. Wu. Variable selection in large margin classifier-based probability estimation with high-dimensional predictors. *Biometrical Journal*, 56(4):594–596, 2014.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

M. Sinistyn. Coordination of price promotions in complementary categories. *Management Science*, 58(11):2076–2094, 2012.

R.J. Slotegraaf and K. Pauwels. The impact of brand equity and innovation on the long-term effectiveness of promotions. *Journal of Marketing Research*, 45(3): 293–306, 2008.

I. Song and P. Chintagunta. Cross-category price effects with aggregate store data. *Management Science*, 52(10):1594–1609, 2006.

I. Song and P. Chintagunta. A discrete-continuous model for multicategory purchase behaviour of households. *Journal of Marketing Research*, 44(4):595–612, 2007.

C.C.A. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing genome-wide association studies: sample size, power, imputation and the choice of genotyping chip. *PLOS Genetics*, 5:e1000477, 2009.

S. Srinivasan, P.T.L. Popkowski Leszczyc, and F.M. Bass. Market share response and competitive interaction: The impact of temporary, evolving and structural changes in prices. *International Journal of Research in Marketing*, 17(4):281–305, 2000.

S. Srinivasan, K. Pauwels, D.M. Hanssens, and M.G. Dekimpe. Do promotions benefit manufacturers, retailers, or both? *Management Science*, 50(5):617–629, 2004.

J. Steenkamp, V. Nijs, D.M. Hanssens, and M.G. Dekimpe. Competitive reactions to advertising and promotion attacks. *Marketing Science*, 24(1):35–54, 2005.

S. Stieglitz and L. Dang-Xuan. Emotions and information diffusion in social media- sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29(4):217–247, 2013.

J.H. Stock and M.W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:1167–1179, 2002.

L. Sun, S. Ji, and J. Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):194–200, 2011.

S. Taskinen, C. Croux, A. Kankainen, E. Ollila, and H. Oja. Canonical analysis based on scatter matrices. *Journal of Multivariate Analysis*, 97:359–384, 2006.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

Y. van Everdingen, D. Fok, and S. Stremersch. Modeling global spillover of new product takeoff. *Journal of Marketing Research*, 46(5):637–652, 2009.

W.N. van Wieringen and M.A. van de Wiel. Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, 65:19–29, 2009.

H.D. Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4:147–166, 1976.

S. Visuri, V. Koivunen, and H. Oja. Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91(2):557–575, 2000.

J. Vuchelen. Consumer sentiment and macroeconomic forecasts. *Journal of Economic Psychology*, 25:493–506, 2004.

S. Waaijenborg, P. Hamer, and A.H. Zwinderman. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(1): Article 3, 2008.

H. Wang and C. Leng. A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 52(12):5277–5286, 2008.

L. Wasserman and K. Roeder. High dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009.

M. Wedel and J. Zhang. Analyzing brand competition across subcategories. *Marketing Science*, 41(4):448–456, 2004.

A. Weinstein. *Handbook of market segmentation: Strategic targeting for business and technology firms, Third Edition.* The Haworth Press, 2013.

B. Wild, M. Eichler, H. C. Friedrich, M. Hartmann, S. Zipfel, and W. Herzog. A graphical vector autoregressive modelling approach to the analysis of electronic diary data. *BMC Medical Research Methodology*, 10(28):1–13, 2010.

I. Wilms and C. Croux. Sparse canonical correlation analysis from a predictive point of view. *Biometrical Journal*, 57(5):834–851, 2015.

D.M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society Series B*, 71:615–636, 2009.

D.M. Witten, R. Tibshirani, and S. Gross. *Package: PMA*, 2011. URL `http://cran.r-project.org/web/packages/PMA/index.html`. R package version 1.0.8.

H.O.A. Wold. *Nonlinear estimation by iterative least square procedures.* Wiley, New York, 1968. URL `http://books.google.be/books?id=hGZ3uAAACAAJ`.

T.T. Wu, Y.F. Chen, and T. Hastie. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

J. Yin and H. Li. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, 5(4):2630–2650, 2011.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 2006.

M. Yuan, V.R. Joseph, and Y. Lin. An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49:430–439, 2007.

X. Zhou, J. Nakajima, and M. West. Bayesian forecasting and portfolio decisions using dynamic sparse factor models. *International Journal of Forecasting*, 30(4): 963–980, 2014.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(5):301–320, 2005.

P.P. Zubcsek and M. Sarvary. Advertising to a social network. *Quantitative Marketing and Economics*, 9(1):71–107, 2011.

Doctoral dissertations of the Faculty of Business and Economics

A list of doctoral dissertations from the Faculty of Business and Economics can be found at the following website:
`http://www.kuleuven.be/doctoraatsverdediging/archief.htm`.